COLORECTAL CANCER - EHR

Alice Brunazzi, 864566

Overview

Introduction

Colorectal cancer (CRC) is among the most common malignancies worldwide and constitutes one of the leading **causes of cancer-related mortality**. In patients with advanced (stage IV) disease, surgical management is particularly complex and the prognosis highly variable, making the identification of **survival-predictive factors** essential for optimizing therapeutic decisions and improving long-term outcomes.

> « In patients with advanced colorectal cancer that underwent surgery, which factors influence overall survival? »

Tai, Ying-Hsuan; Chang, Wen-Kuei; Wu, Hsiang-Ling; Chan, Min-Ya; Chen, Hsiu-Hsi; Chang, Kuang-Yi (2018). 4 Dataset.. PLOS ONE. Dataset. https://doi.org/10.1371/journal.pone.0200893.s001

Dataset

DEMOGRAPHICS

1. GENDER: binary, 1 male and 2 female
2. AGE: mean 65, minumun 18, maximum 95
3. ASA3 indicator

PROGNOSIS

1. Tumor Location
2. CEA value, tumoral marker
3. Cell Differenciation, liver metastasis,
4. Mucinous type, Signet Ring, Lymphovascular invasion

COMORBIDITIES

1. Diabetes, kidney disease
2. Coronary artery disease, heart failure

SURGERY INFO

1. Time under Anestesia, mean 342 minutes, min 50 max 960
2. Units of blod trasferred during operation

TREATMENT and RESULTS

1. Death and Progression
2. Time between the operation and the Reoccurence or death
3. Radiotherapy, Chemioteraphy, Neoadjuvant therapy

Dataset

DEMOGRAPHICS

1. GENDER: binary, 1 male and 2 female
2. AGE: mean 65, minumun 18, maximum 95
3. ASA3 indicator

Both ASA and ASA3 (corr. 90%) were provided, to choose which was the best choice to keep:

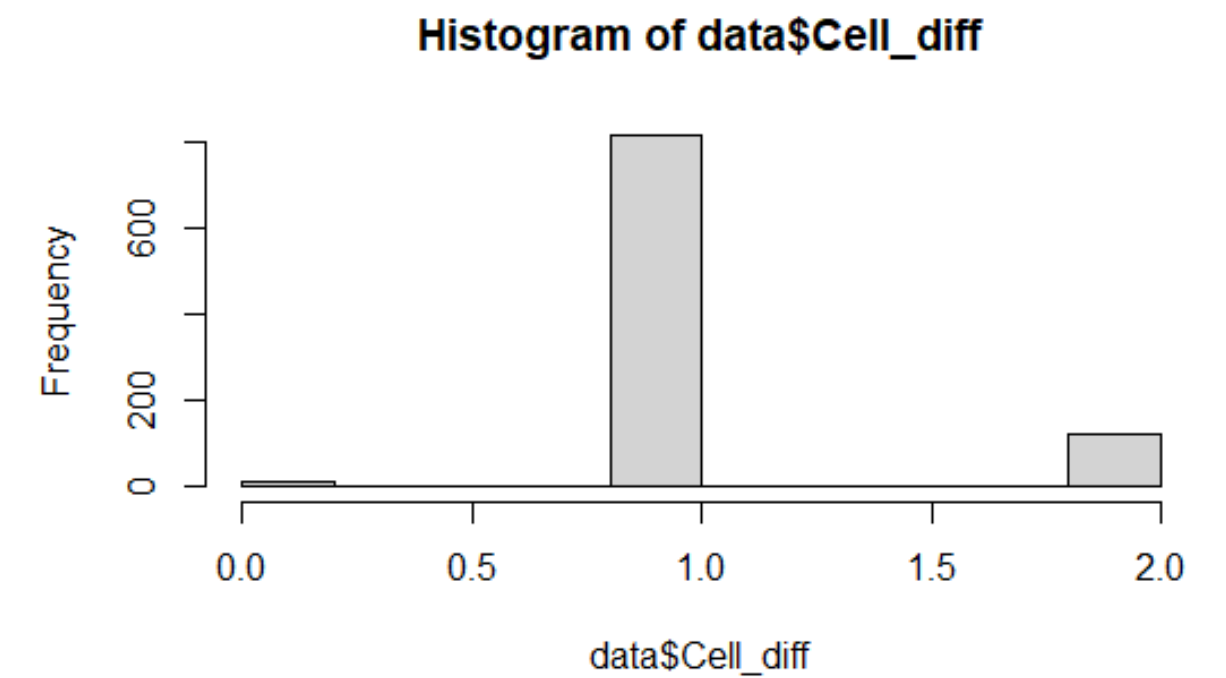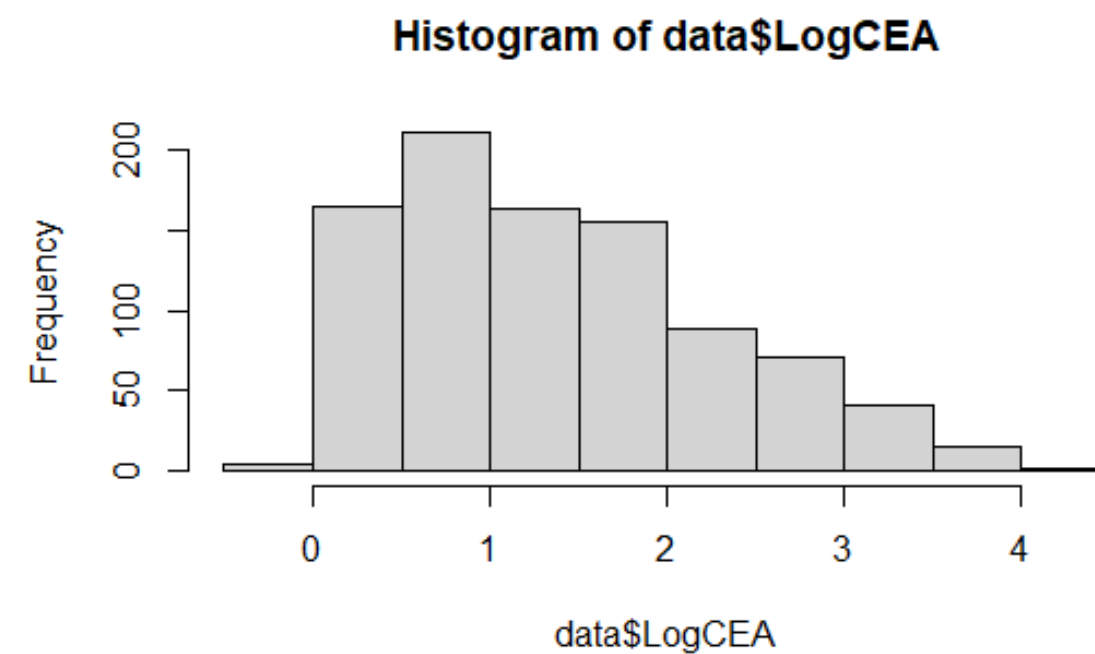|	df <dbl>	AIC <dbl>
model1	2	1366.750
model2	2	1363.609

Dataset

PROGNOSIS

1. Tumor Location
2. CEA value, tumoral marker
3. Cell Differenciation, liver metastasis,
4. Mucinous type, Cellular differenciation, Signet Ring, Lymphovascular invasion

Perineural and Cell Differentiation contained 13 NULLs, eliminated. LogCEA 15 imputation with value 1.264

Dataset

COMORBIDITIES

1. Diabetes, kidney disease
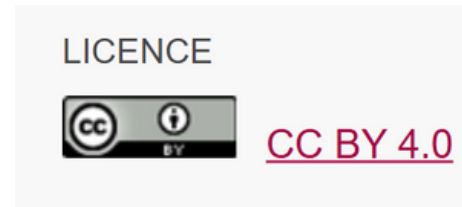2. Coronary artery disease, heart failure

```
Variabile: DM
[1] 0 1

Variabile: CAD
[1] 0 1

Variabile: HF
[1] 0 1

Variabile: CVA
[1] 0 1

Variabile: CKD
[1] 0 1
```

Dataset

SURGERY INFO

1. Time under Anestesia, mean 342 minutes, min 50 max 960
2. Units of blod trasferred during operation

TREATMENT and RESULTS

1. Death and Progression
2. Time between the operation and the Reoccurence or death
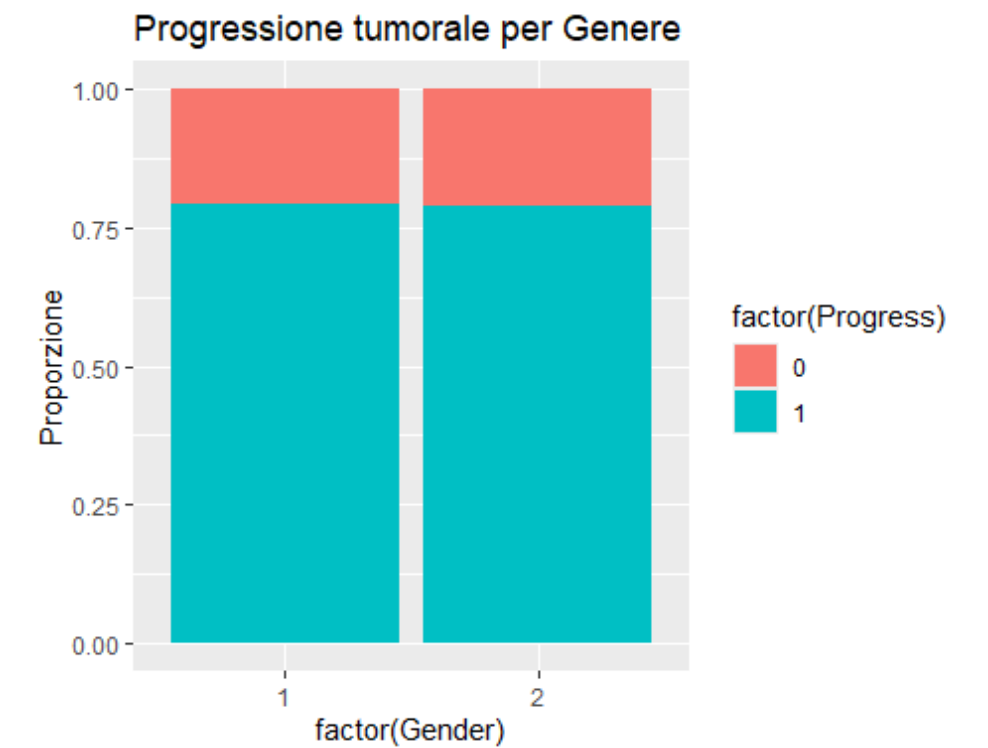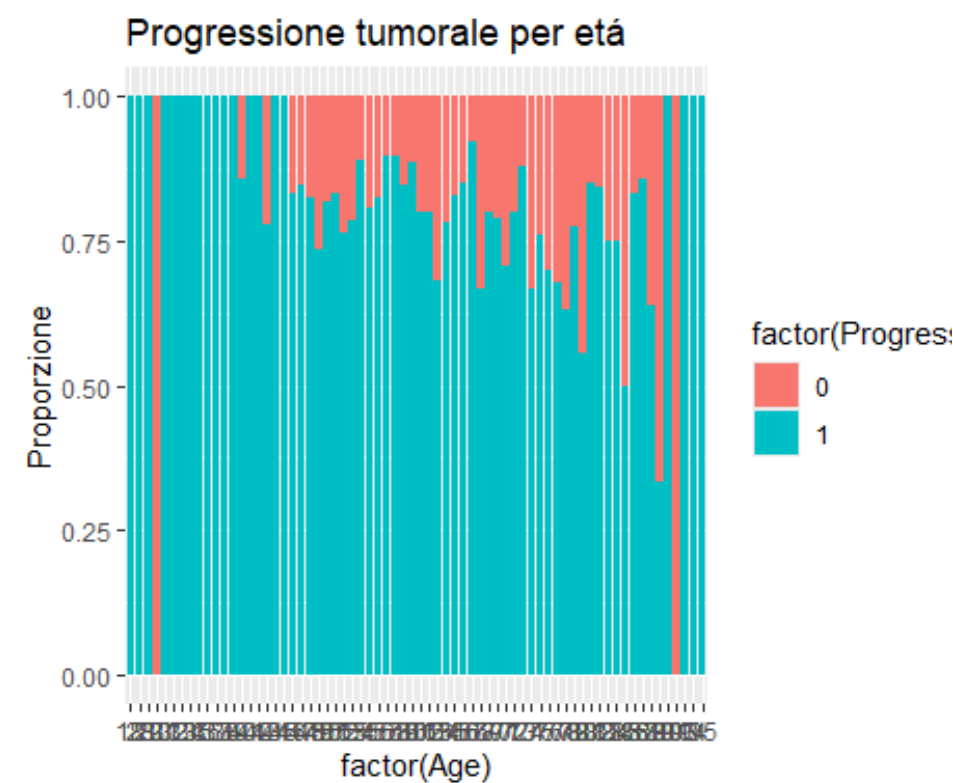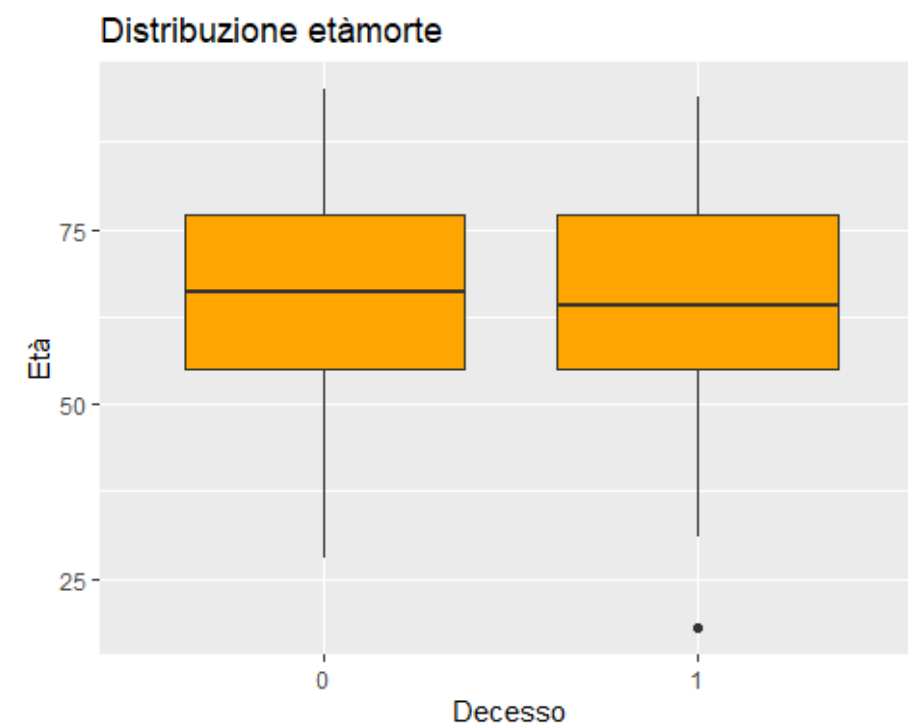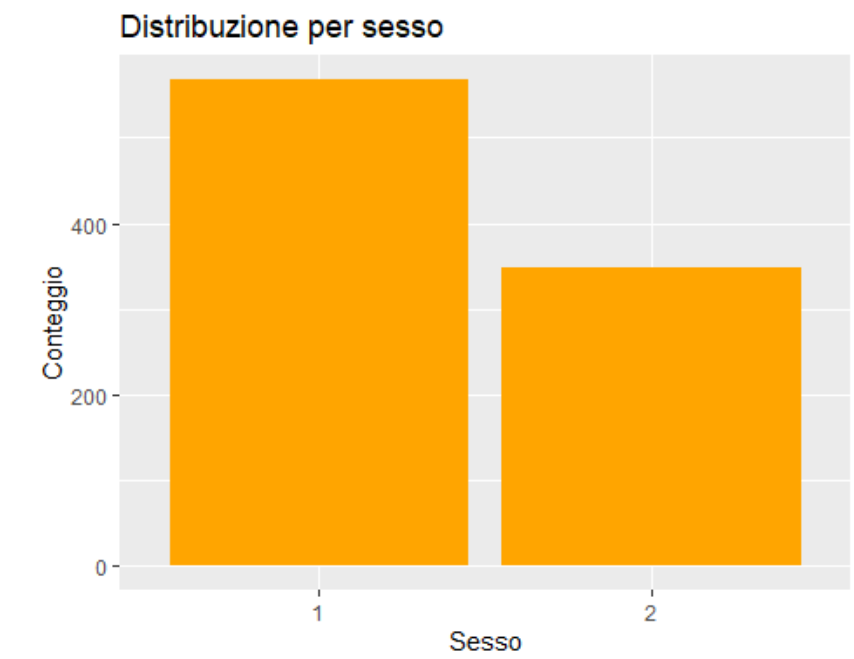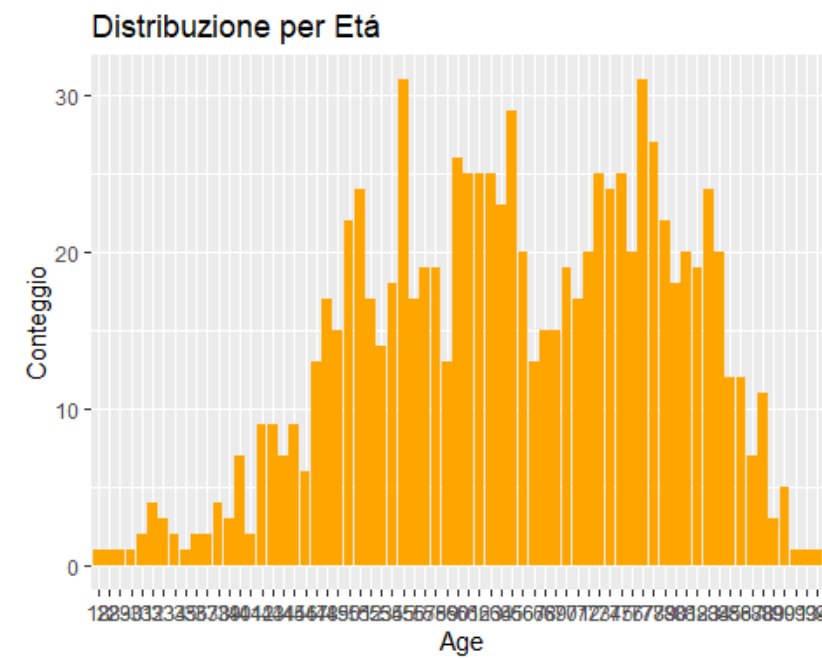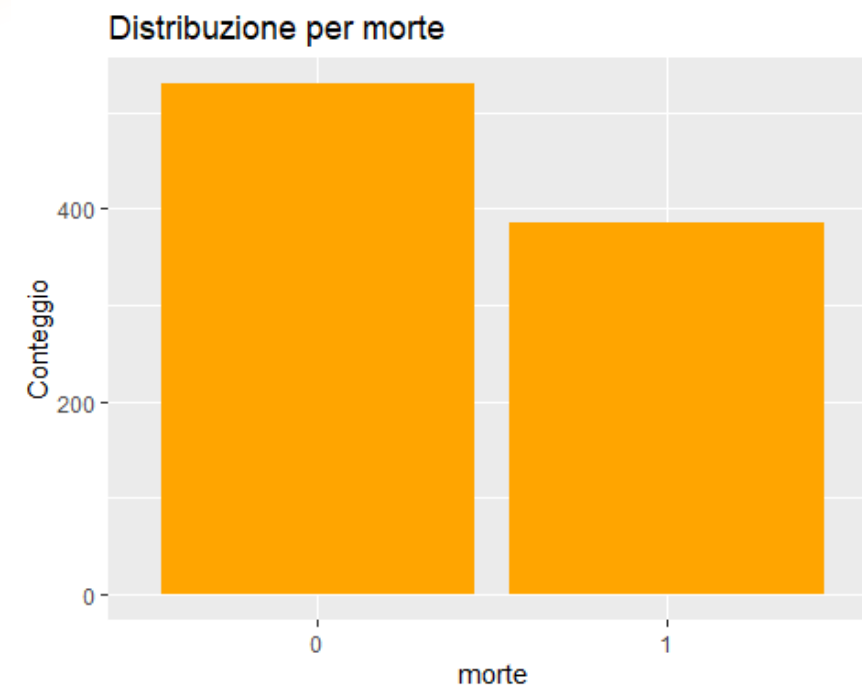3. Radiotherapy, Chemioteraphy, Neoadjuvant therapy

916 RECORDS
28 NUMERIC VARIABLES
6 NON- BINARY VARIABLES

DATA VISUALIZATION

DATA VISUALIZATION

AnesTime vs Survival Time by Death
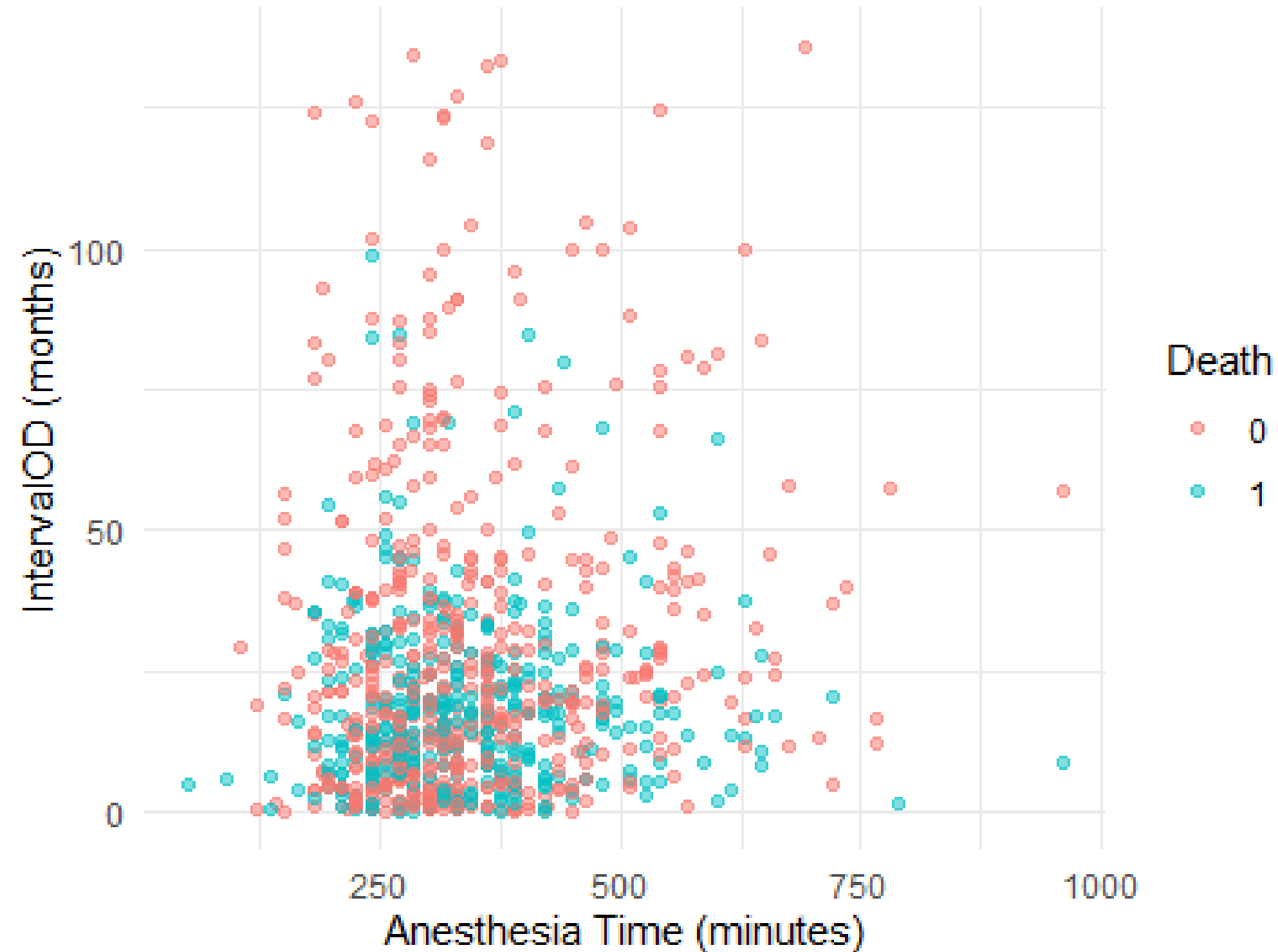
EXPLORATORY DATA ANALYSIS – CORRELATION ANALYSIS
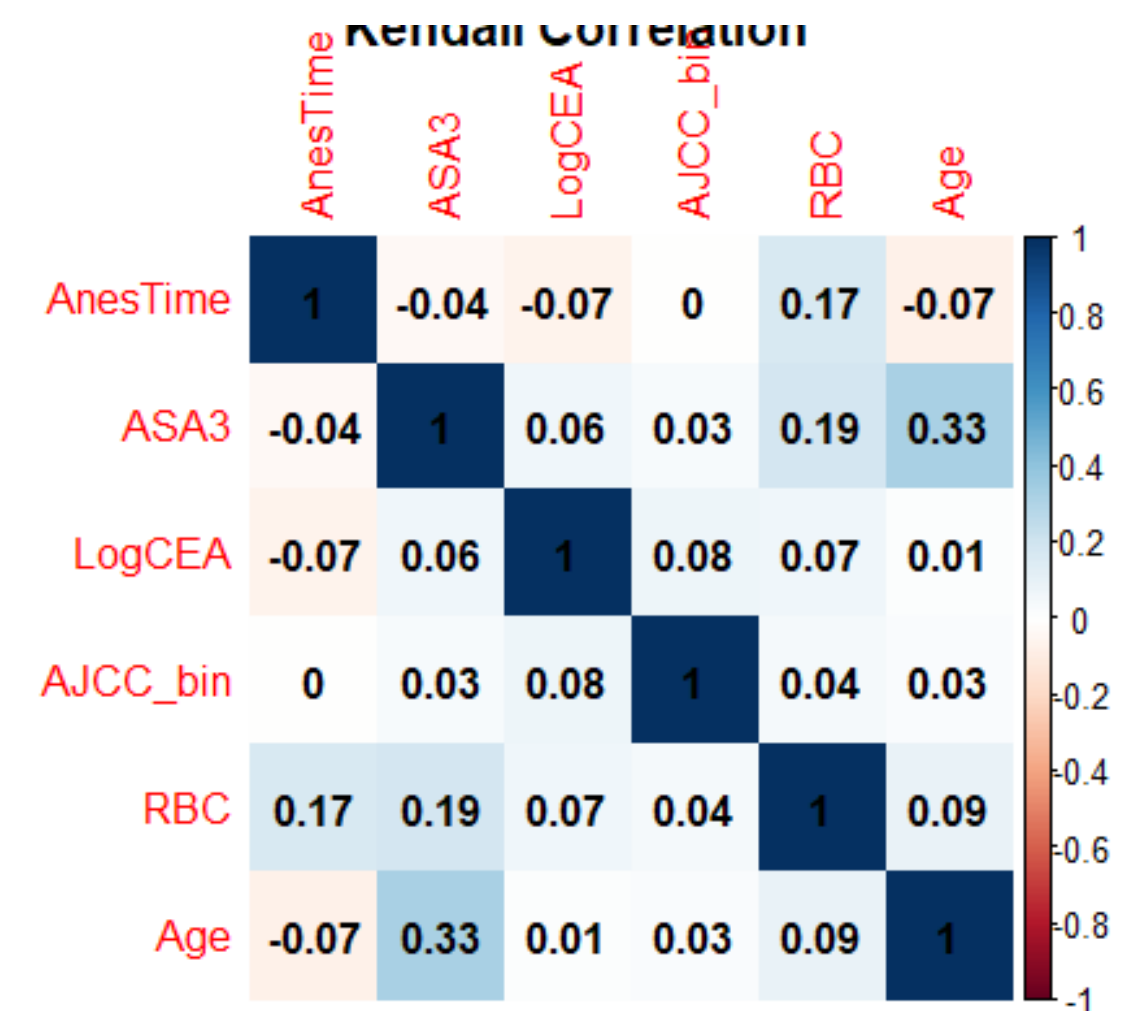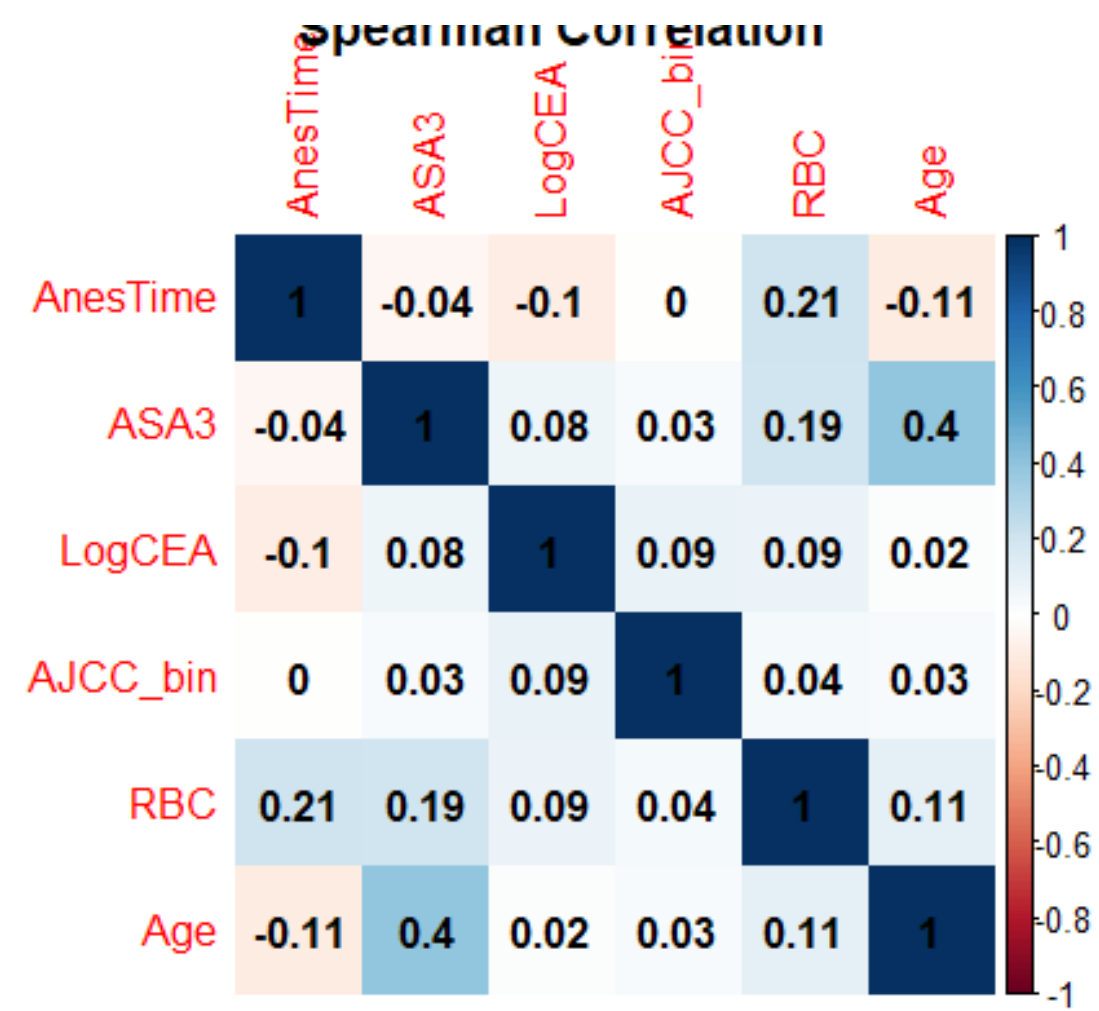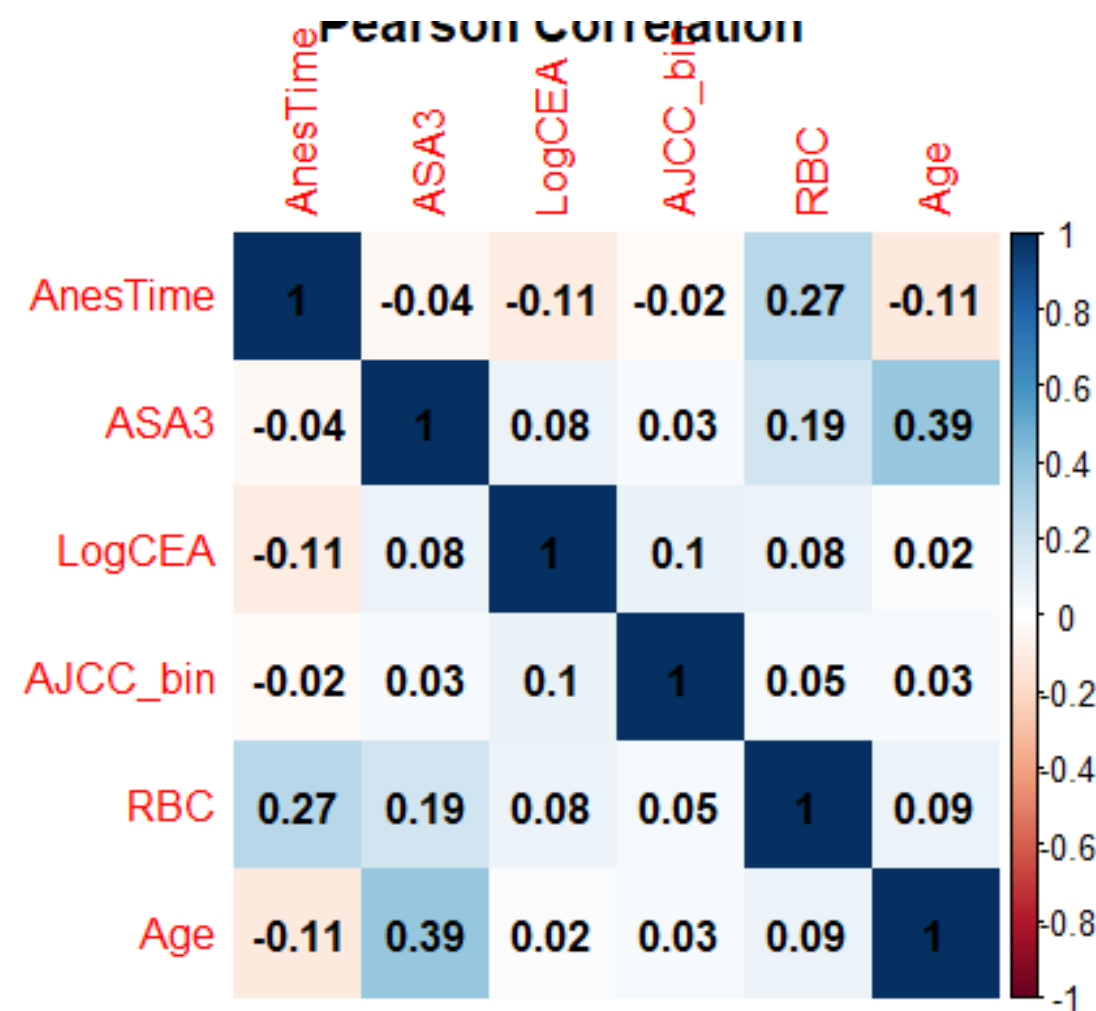
EDA – DIMENSIONALITY REDUCTION
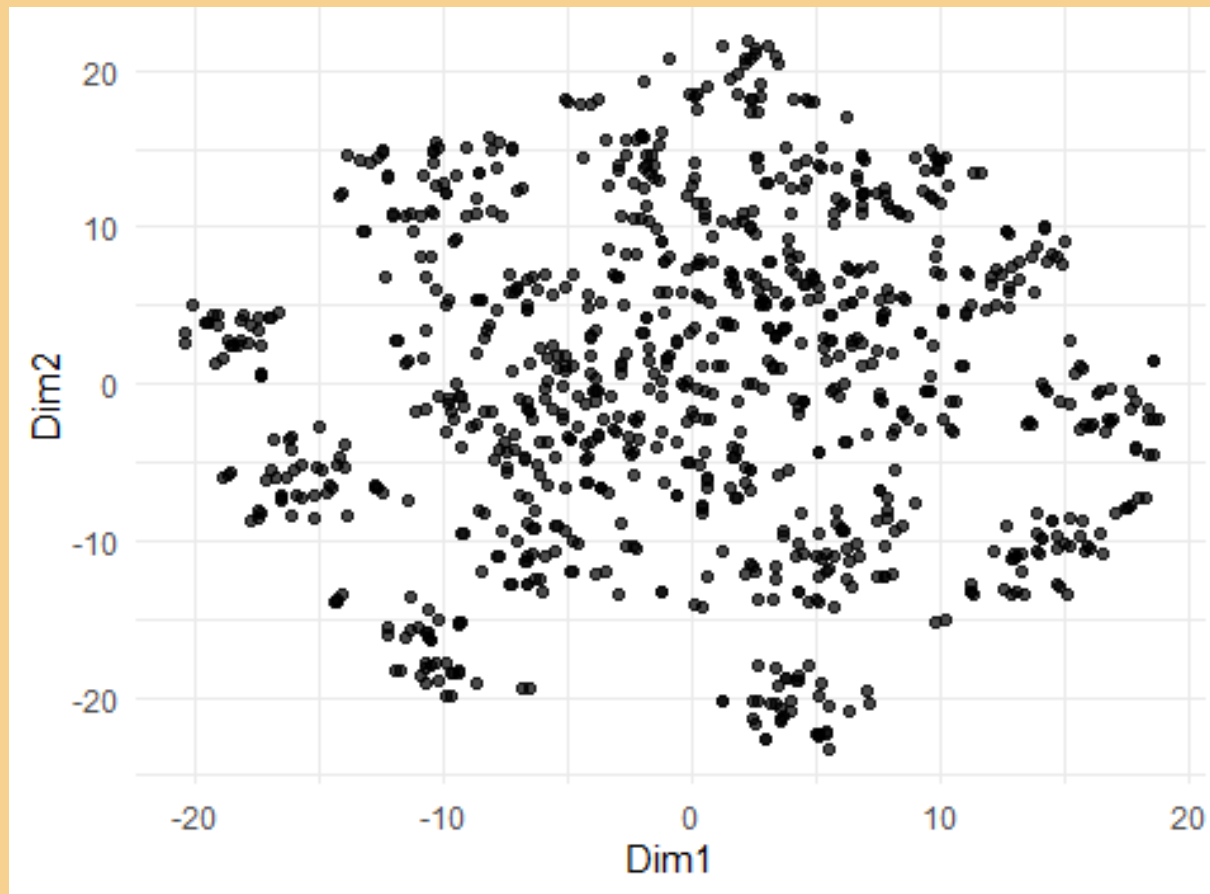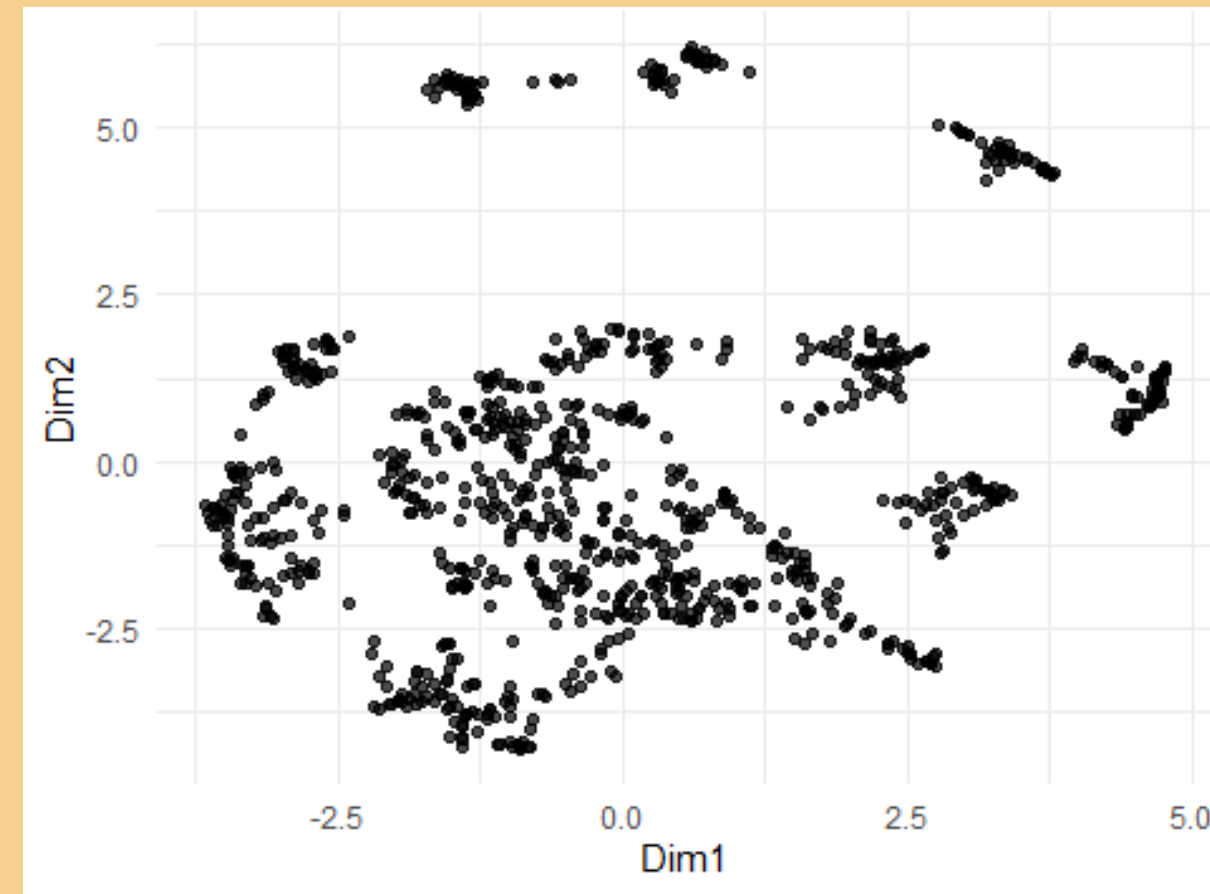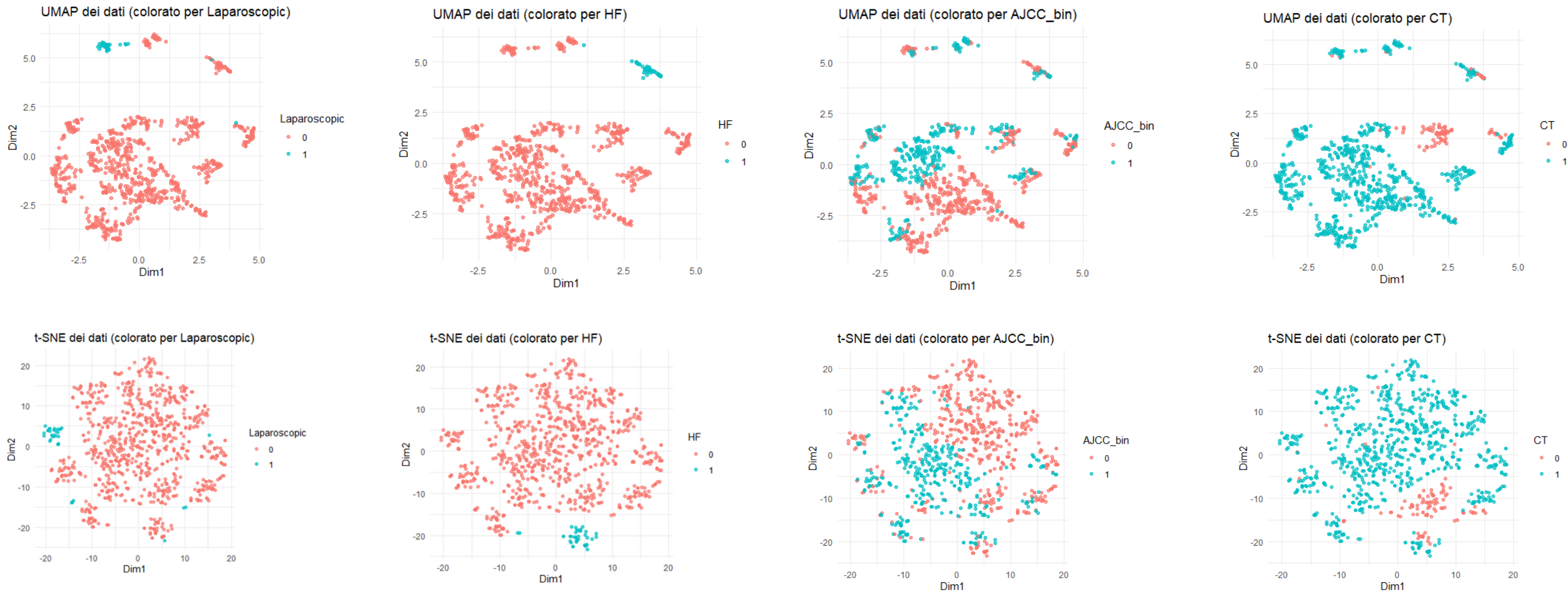
UMAP dei dati (colorato per Laparoscopic) · UMAP dei dati (colorato per HF) · UMAP dei dati (colorato per AJCC_bin) · UMAP dei dati (colorato per CT)

t-SNE dei dati (colorato per Laparoscopic) · t-SNE dei dati (colorato per HF) · t-SNE dei dati (colorato per AJCC_bin) · t-SNE dei dati (colorato per CT)

DIMENSION 1: COMORBIDITIES, CAD, CVA, HF, ASA3 AND CKD. LAPAROSCOPIC SURGERY, TUMOR TREATED WITH CHEMIOTERAPY AND RADIOTERAPHY.

DIMENSION 2: SIGNET RING, MUCIN TYPE, AJCC 4b, CELLULAR DIFFERENCIATION AND LYMPHOVASCULAR INVASION. MORE ADVANCED TREATMENT, NACTRT, METASTASIS

DIMENSION 1: COMORBIDITIES, CAD, CVA, HF, ASA3 AND CKD. LAPAROSCOPIC SURGERY, TUMOR TREATED WITH CHEMIOTERAPY AND RADIOTERAPHY.

|	Variabile <chr>	p_value <dbl>	t_stat <dbl>	mean_group0 <dbl>	mean_group1 <dbl>
t4	HF	2.767488e-71	-35.14373960	-0.150515269	3.296284383
t5	CVA	6.328611e-54	-34.99370912	-0.261232283	4.253650200
t3	CAD	1.996843e-51	-24.84879965	-0.233714234	2.914553971
t17	RT	4.870603e-40	18.83061115	0.329947448	-2.754051147
t16	CT	8.900871e-34	15.99985761	2.249836110	-0.257280528
t19	Progress	2.866483e-24	11.05083347	1.336401681	-0.352072719
t1	ASA3	1.413377e-17	-8.80017433	-0.477899483	0.790958275
t18	NACTRT	1.116663e-10	6.75148089	0.167023580	-0.902861728
t20	AJCC_bin	2.882267e-08	5.60154170	0.323735354	-0.446502527
t7	Laparoscopic	1.317055e-06	5.54393303	0.045285016	-1.075825115
t12	Mucin_TYPE	5.987378e-06	4.85298841	0.095623467	-1.173812705
t6	CKD	8.493830e-06	-4.60854581	-0.135804692	0.875553824
t10	Liver_Only	1.692891e-05	-4.32830692	-0.233175115	0.365111735
t13	SignetRING	1.200959e-04	-4.09967412	-0.029089898	0.620820988
t	Gender	6.193866e-04	3.43743663	NaN	0.183445054
t2	DM	3.601744e-03	-2.93539612	-0.107274975	0.407195585
t9	EA	2.309398e-02	2.28878277	0.072040655	-0.364974179
t8	TumorLOC	9.187436e-02	-1.68887845	-0.080547297	0.182019692
t14	Lymphovascularinvasion	7.963385e-01	-0.25816457	-0.018446510	0.017352226
t15	perineural	8.507105e-01	0.18833281	0.006263335	-0.021587221
t11	Cell_diff	9.226320e-01	-0.09727283	NaN	-0.002413919

21 rows

DIMENSION 2: SIGNET RING, MUCIN TYPE, AJCC 4b, CELLULAR DIFFERENCIATION AND LYMPHOVASCULAR INVASION. MORE ADVANCED TREATMENT, NACTRT, METASTASIS

|	Variabile <chr>	p_value <dbl>	t_stat <dbl>	mean_group0 <dbl>	mean_group1 <dbl>
t4	HF	2.602859e-187	-52.4326789	-0.211624838	4.63458395
t7	Laparoscopic	3.441017e-57	-42.9959622	-0.231252888	5.49381862
t13	SignetRING	4.470704e-27	-22.4893977	-0.259113932	5.52987050
t20	AJCC_bin	1.151128e-25	-10.8171953	-0.709936232	0.97915880
t16	CT	5.394443e-22	11.3304209	1.875427254	-0.21446492
t10	Liver_Only	2.346798e-19	9.2525237	0.590846182	-0.92516251
t18	NACTRT	1.082772e-17	9.5030323	0.349311463	-1.88823609
t12	Mucin_TYPE	1.135163e-14	-9.4702453	-0.214907894	2.63807227
t14	Lymphovascularinvasion	4.234548e-14	-7.6755189	-0.633389404	0.59581546
t11	Cell_diff	1.616003e-12	-7.7310902	-0.253874699	1.73372551
t5	CVA	3.906720e-10	-7.1793838	-0.085290241	1.38878260
t1	ASA3	1.091821e-07	-5.3599393	-0.330081630	0.54630902
t15	perineural	4.035383e-06	-4.6970605	-0.224869993	0.77503735
t8	TumorLOC	2.351017e-04	-3.7045536	-0.207809958	0.46960613
t6	CKD	8.605580e-04	-3.3928952	-0.103636951	0.66816343
t3	CAD	1.648666e-03	-3.2589744	-0.073245440	0.91341372
t17	RT	1.222044e-02	2.5367954	0.055363879	-0.46211891
t19	Progress	2.777496e-02	2.2118533	0.370655547	-0.09764856
t2	DM	2.487374e-01	-1.1556047	-0.047854292	0.18164587
t9	EA	4.059697e-01	0.8325073	0.027991509	-0.14181129
t	Gender	9.121195e-01	0.1104014	0.007026667	-0.01146881

CLUSTERING
UMAP + KMEANS
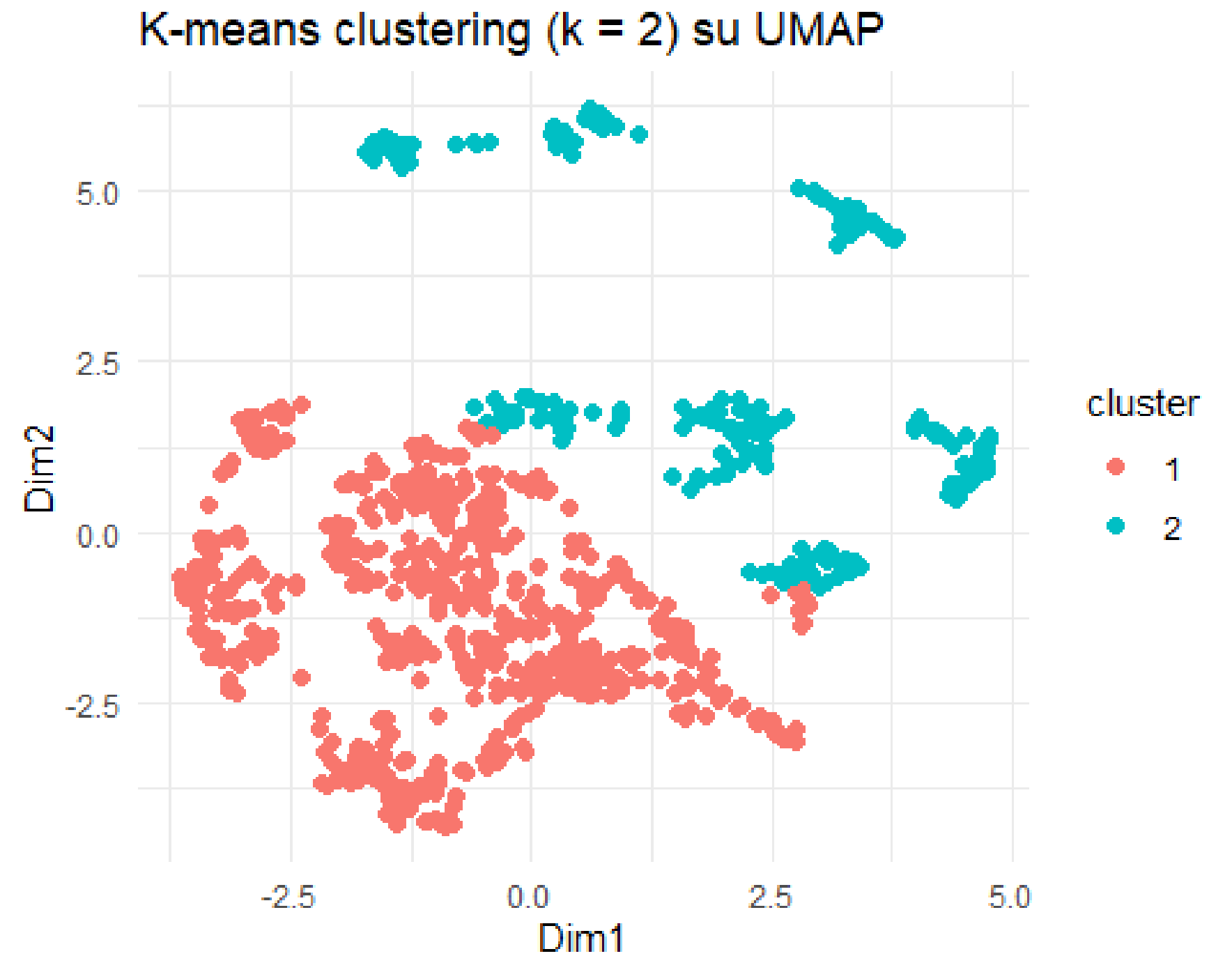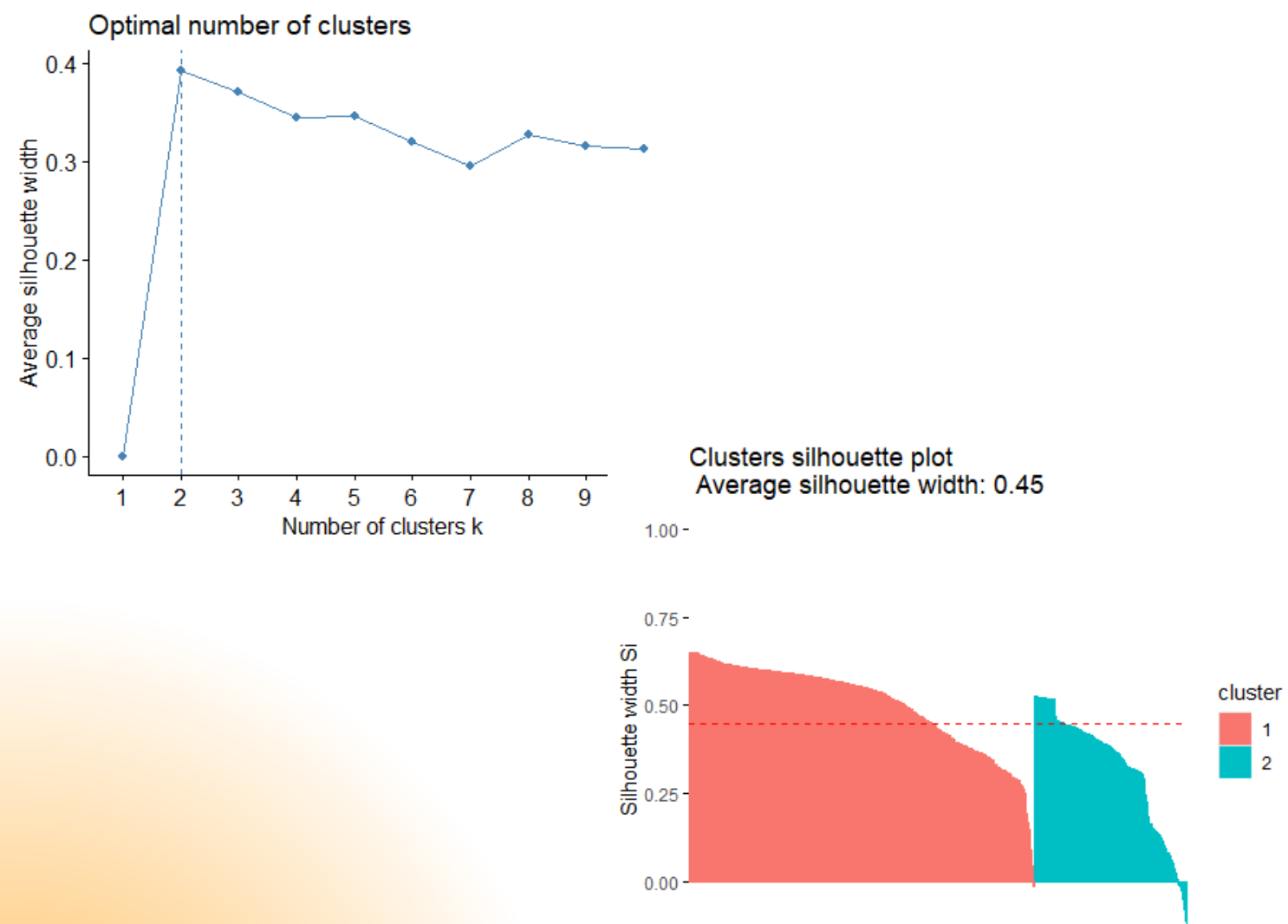
CLUSTERING

Cluster sui primi 18 componenti principali

Cluster plot
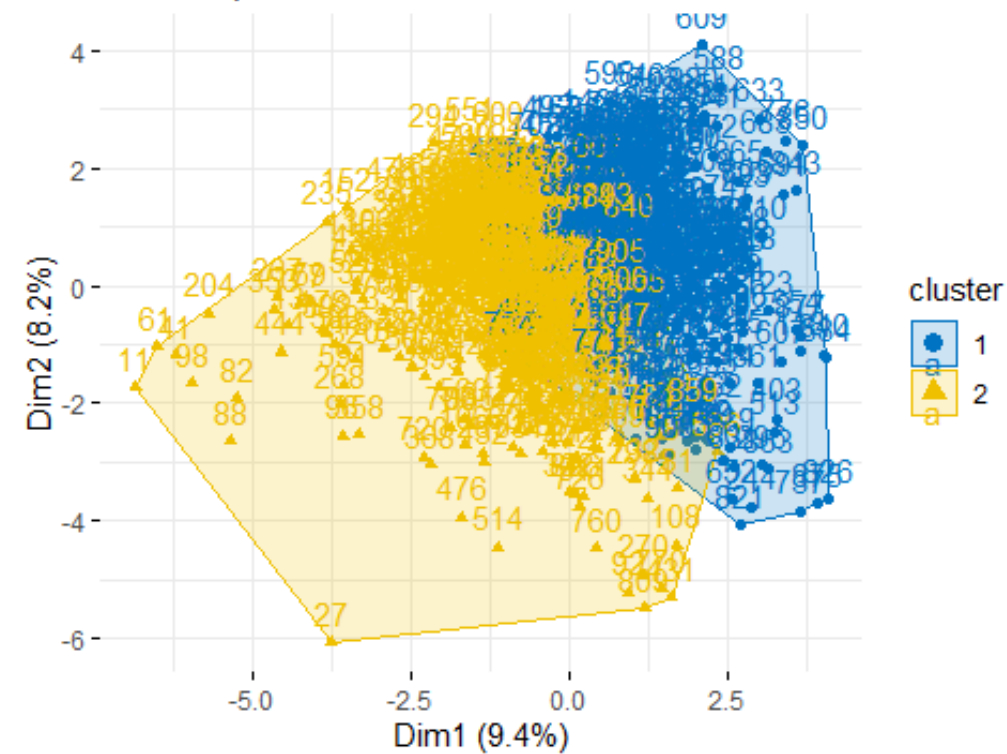
Cluster (da dati originali) visualizzati su UMAP
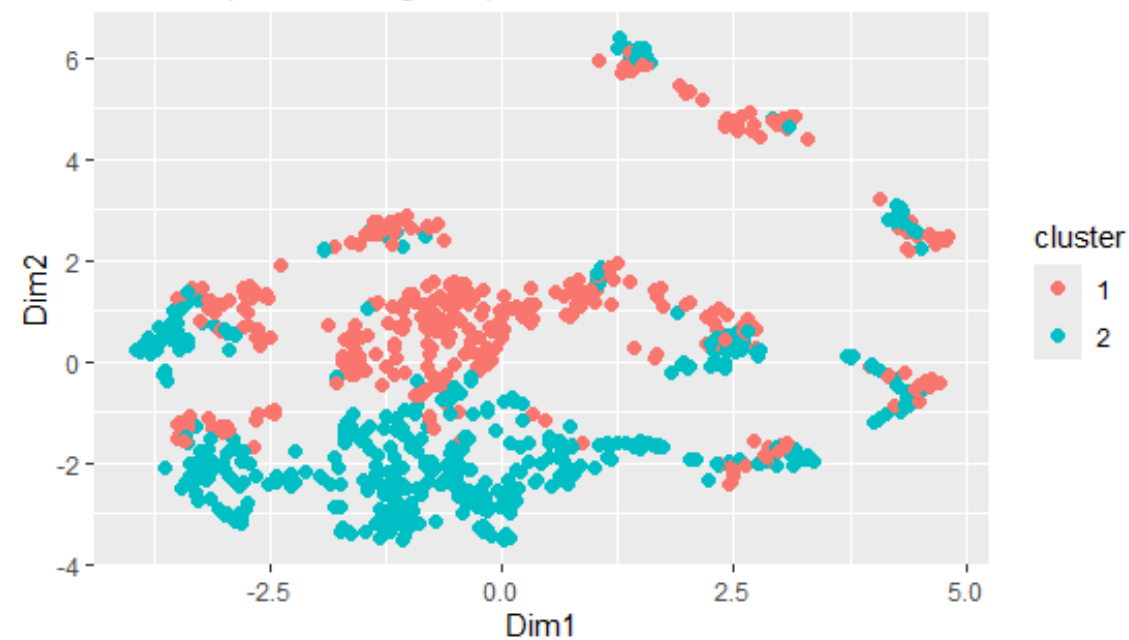
Cluster Dendrogram

CLUSTERING

Cluster comparison (all variables scaled as % of max)
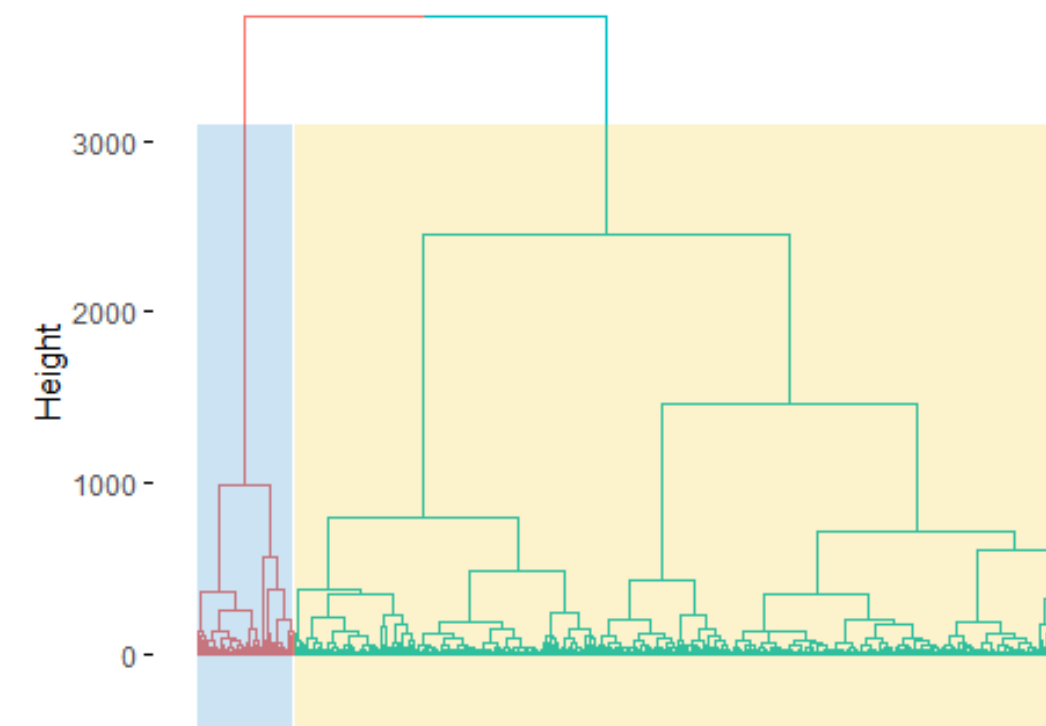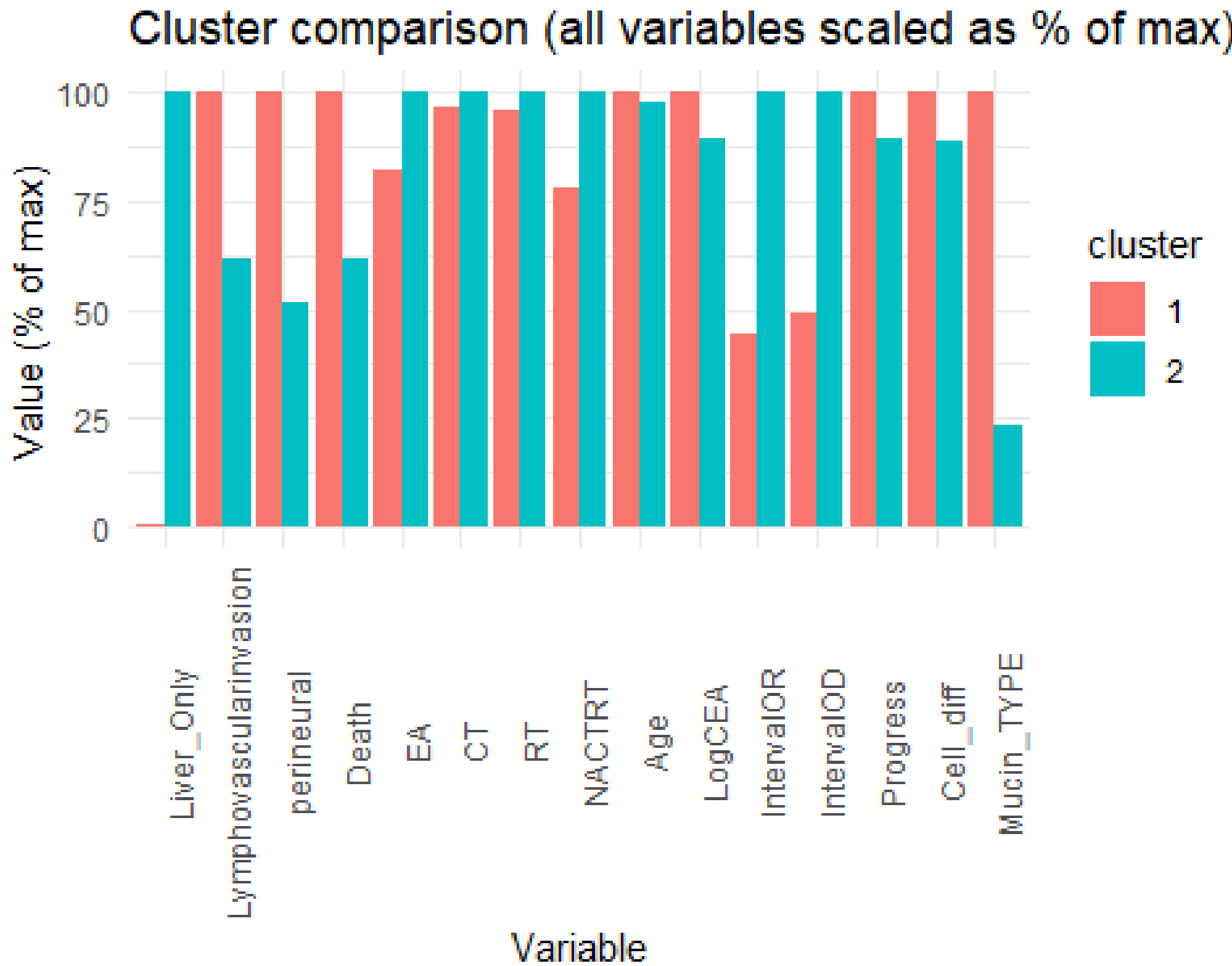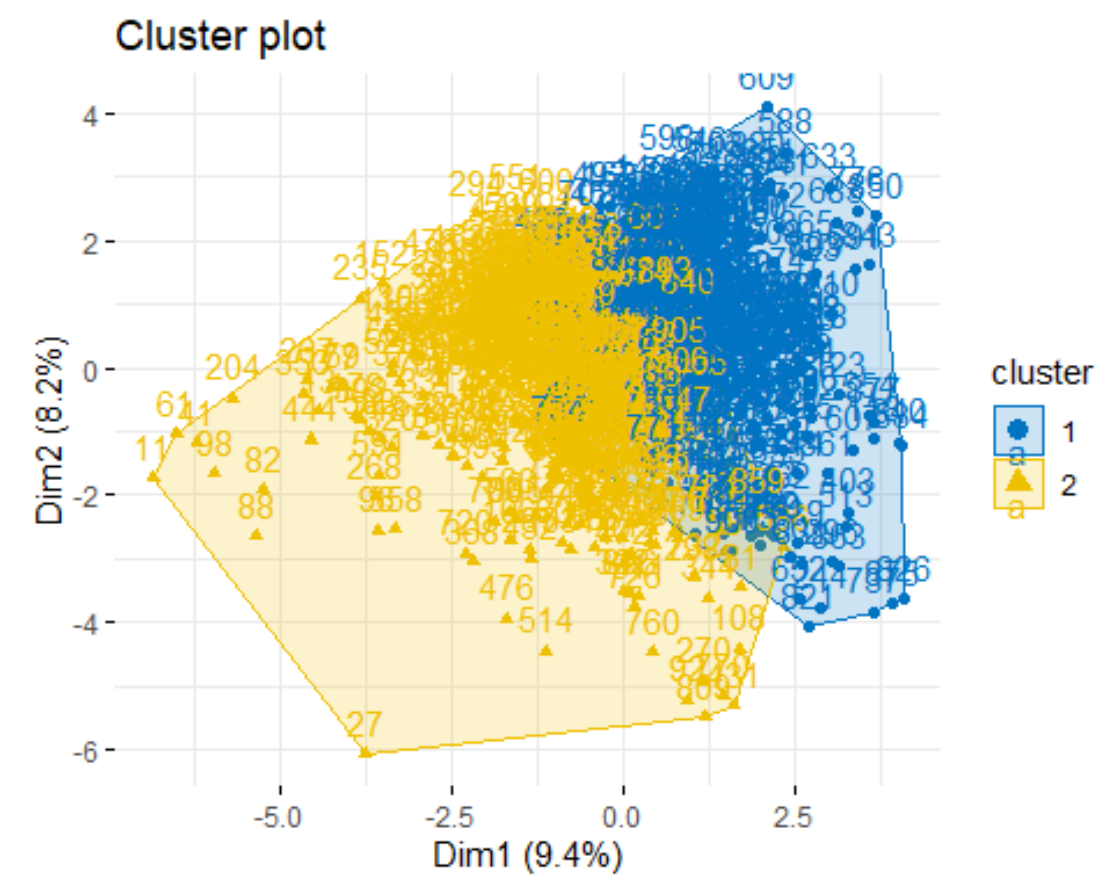
CLUSTERING – KMEANS & HAC

INTERNAL MEASURES

AVERAGE SILHOUETTE (near 1)
HAC: 0.094
KMEANS: 0.069
DAVIES BOULDIN (near 0)
HAC: 3.98
KMEANS: 3.78

HAC had 814 records in cluster 1 and 102 in cluster 2. Kmeans created cluster 1 with 420 records and cluster 2 with 496

ODDS RATIO - GLM

Odds Ratio con Intervallo di Confidenza (95%)

KAPLAN MEIER CURVES

KAPLAN MEIER CURVES

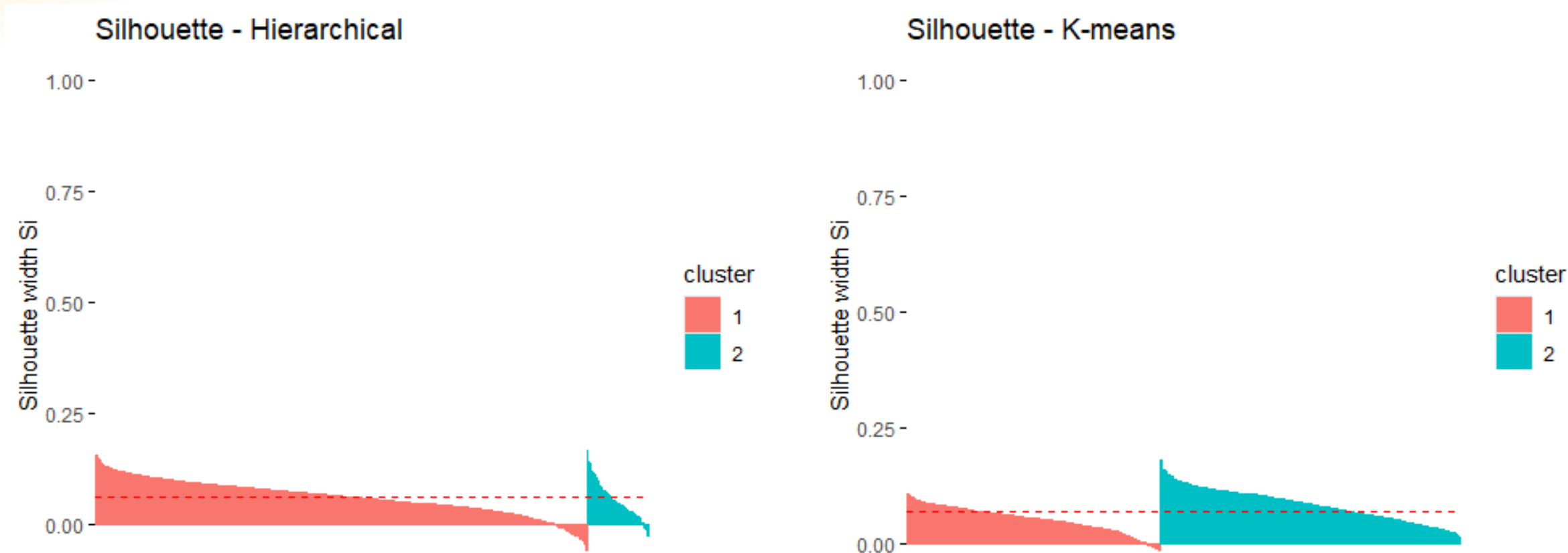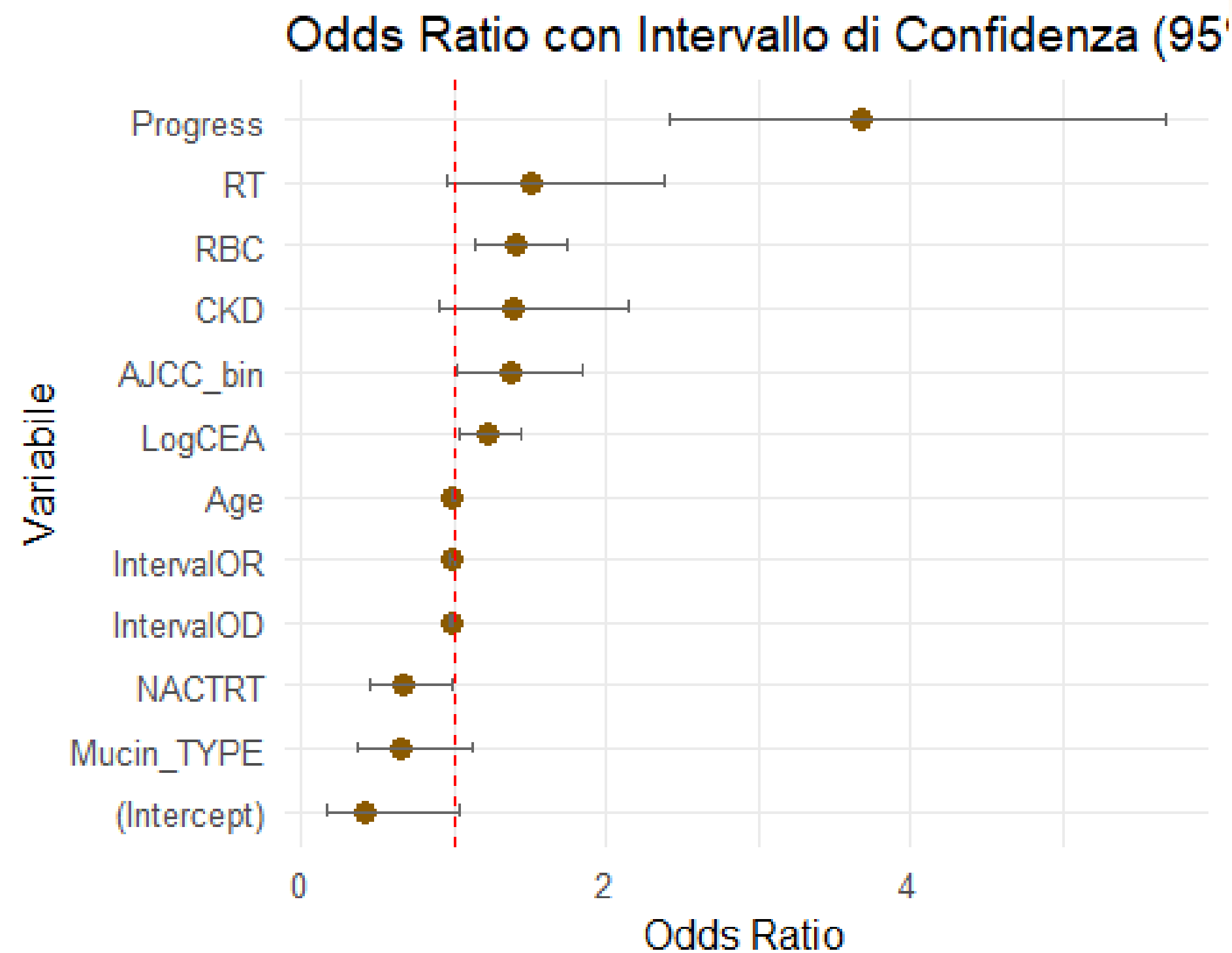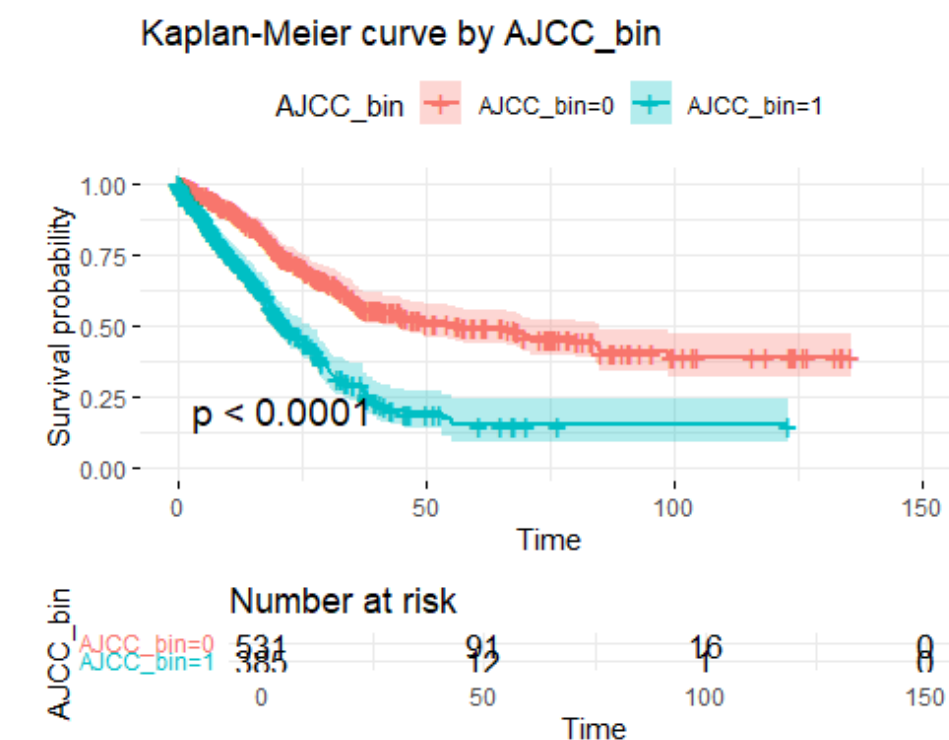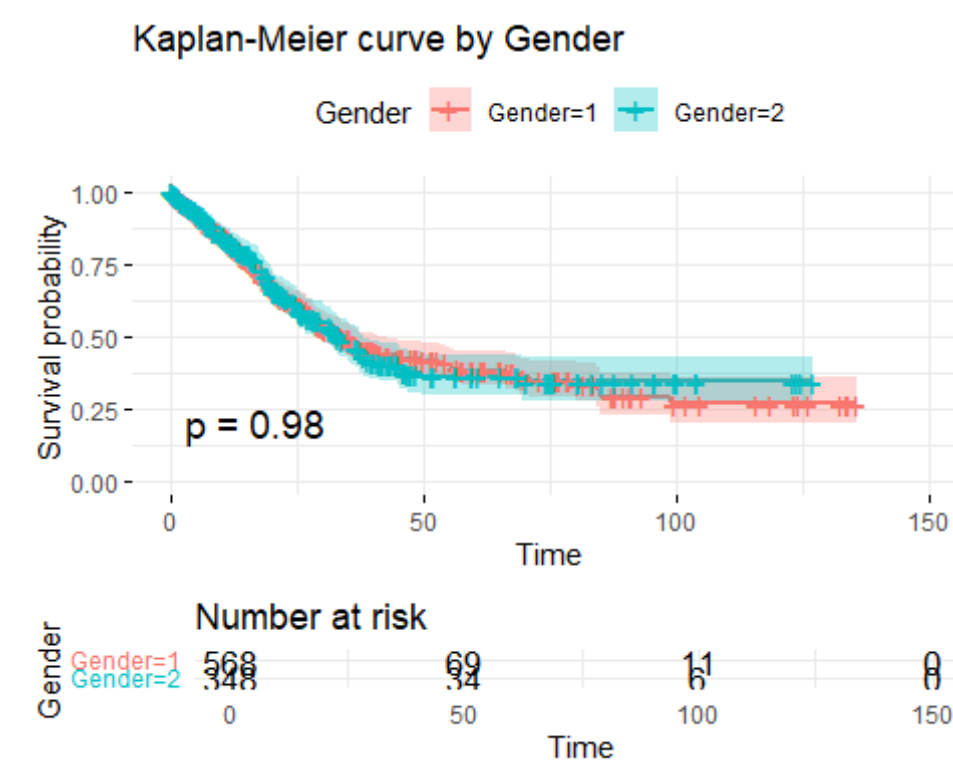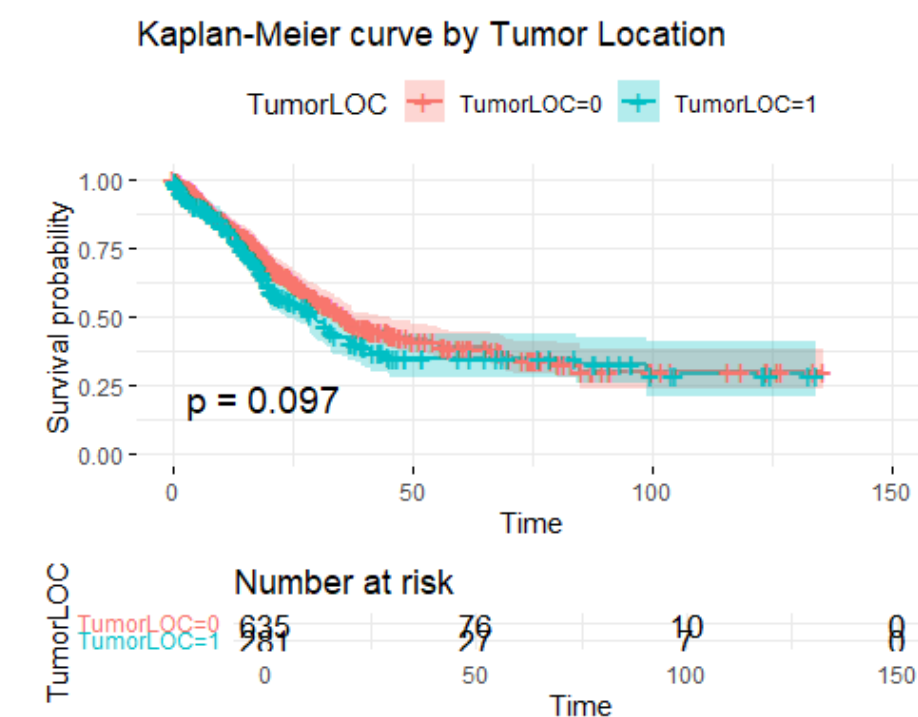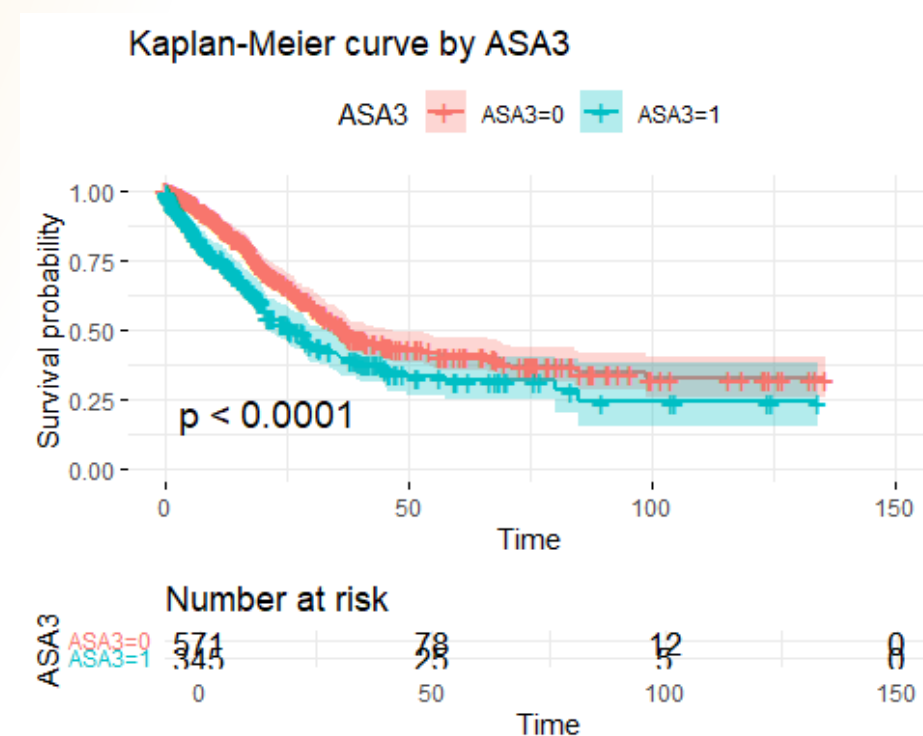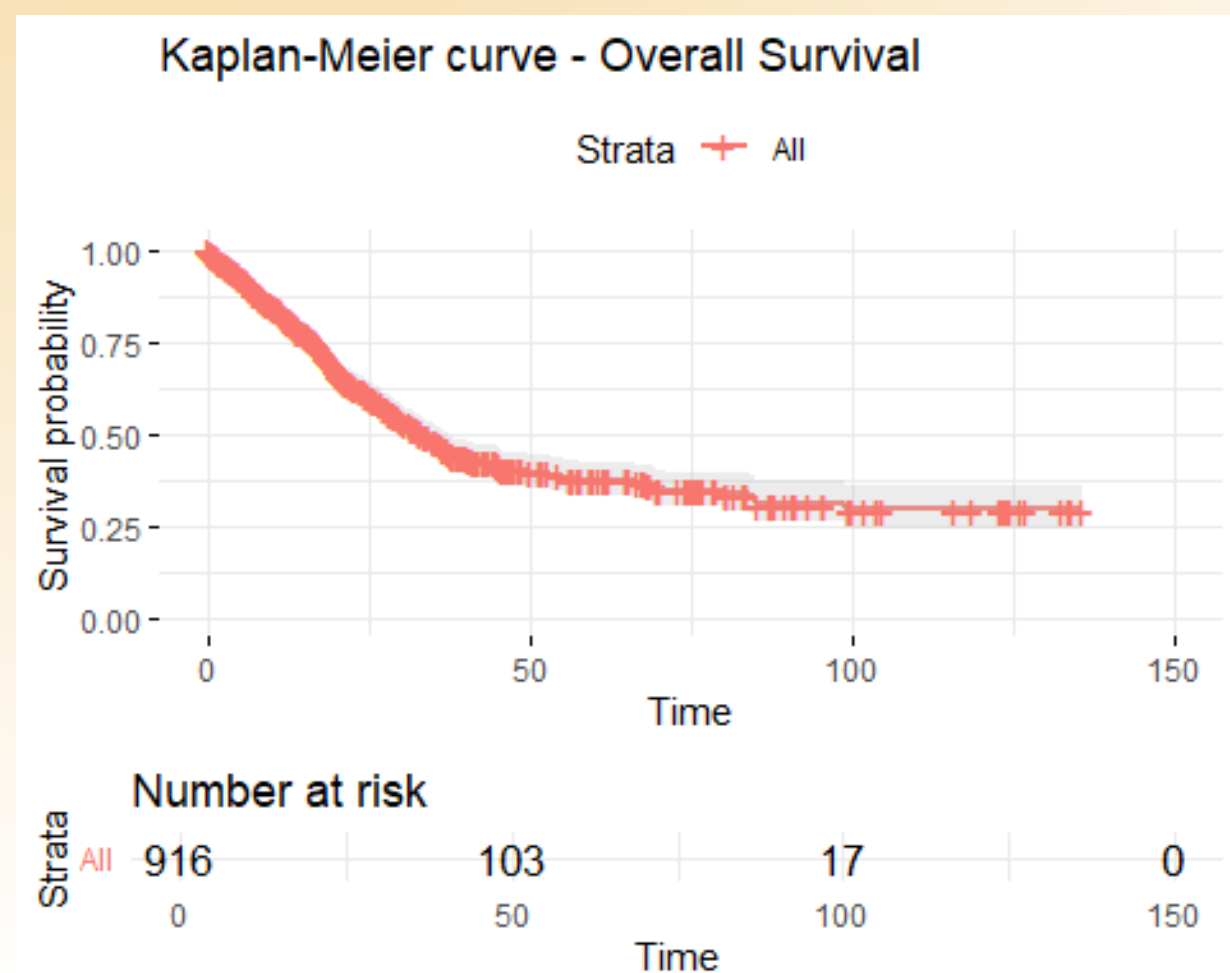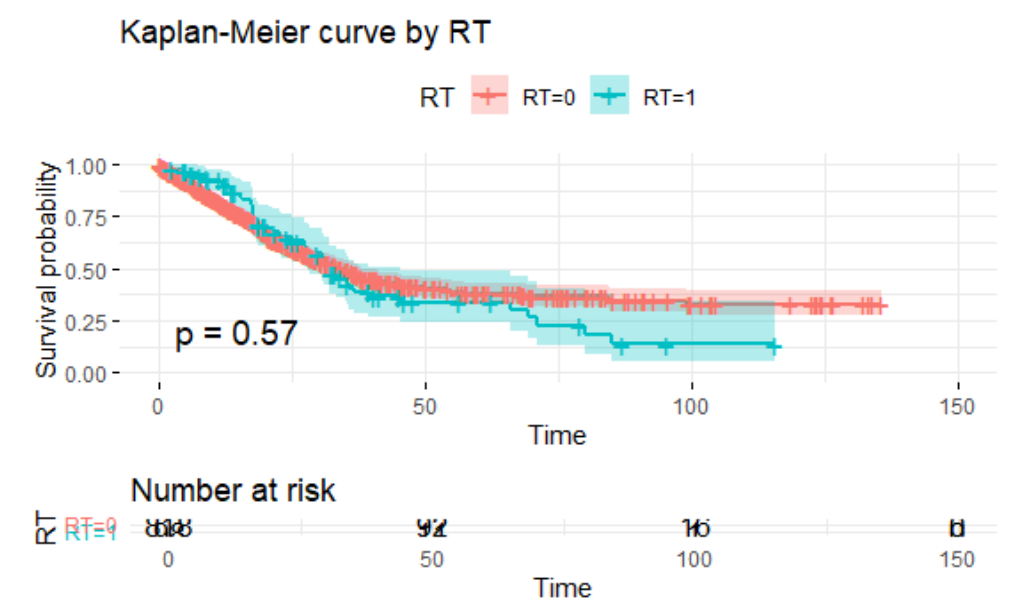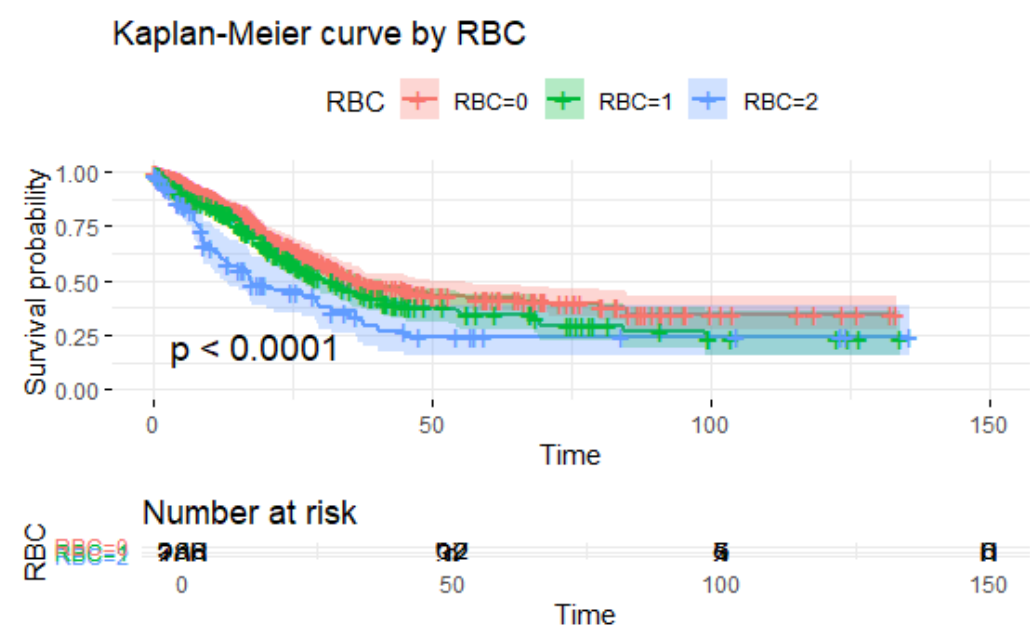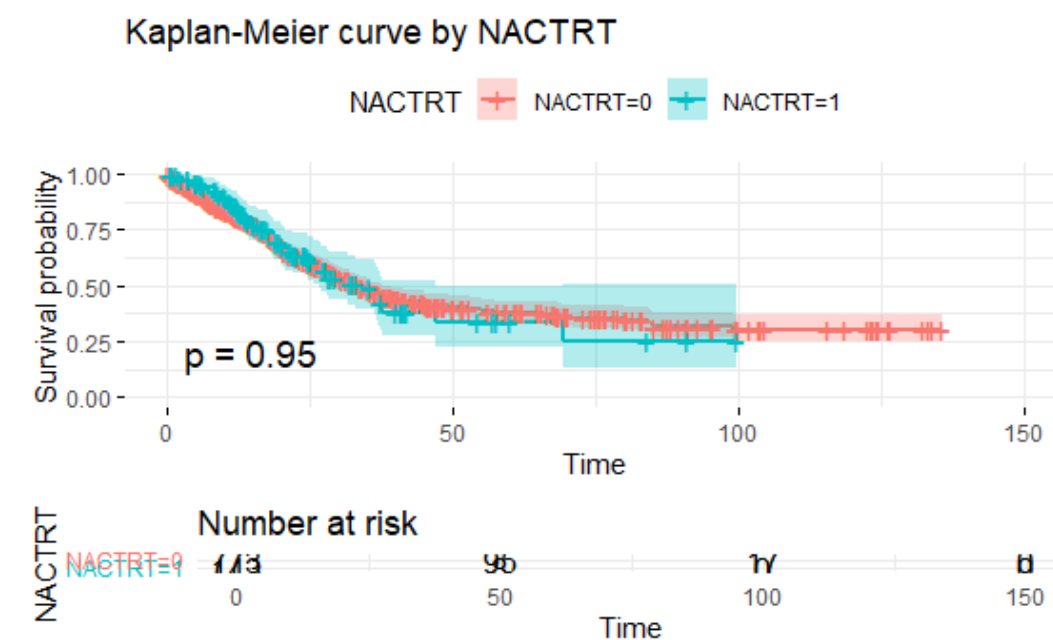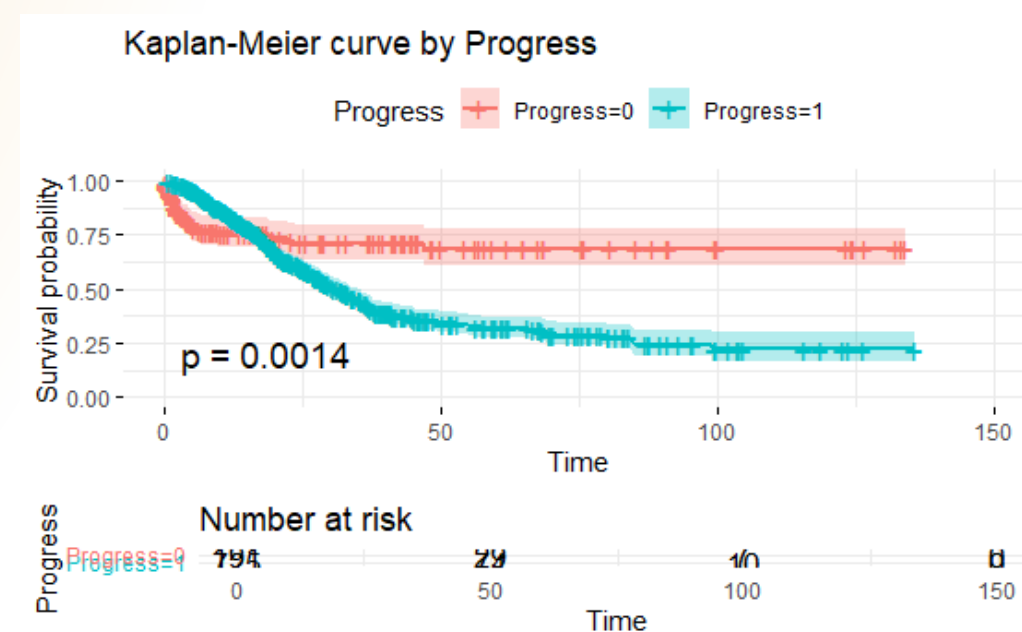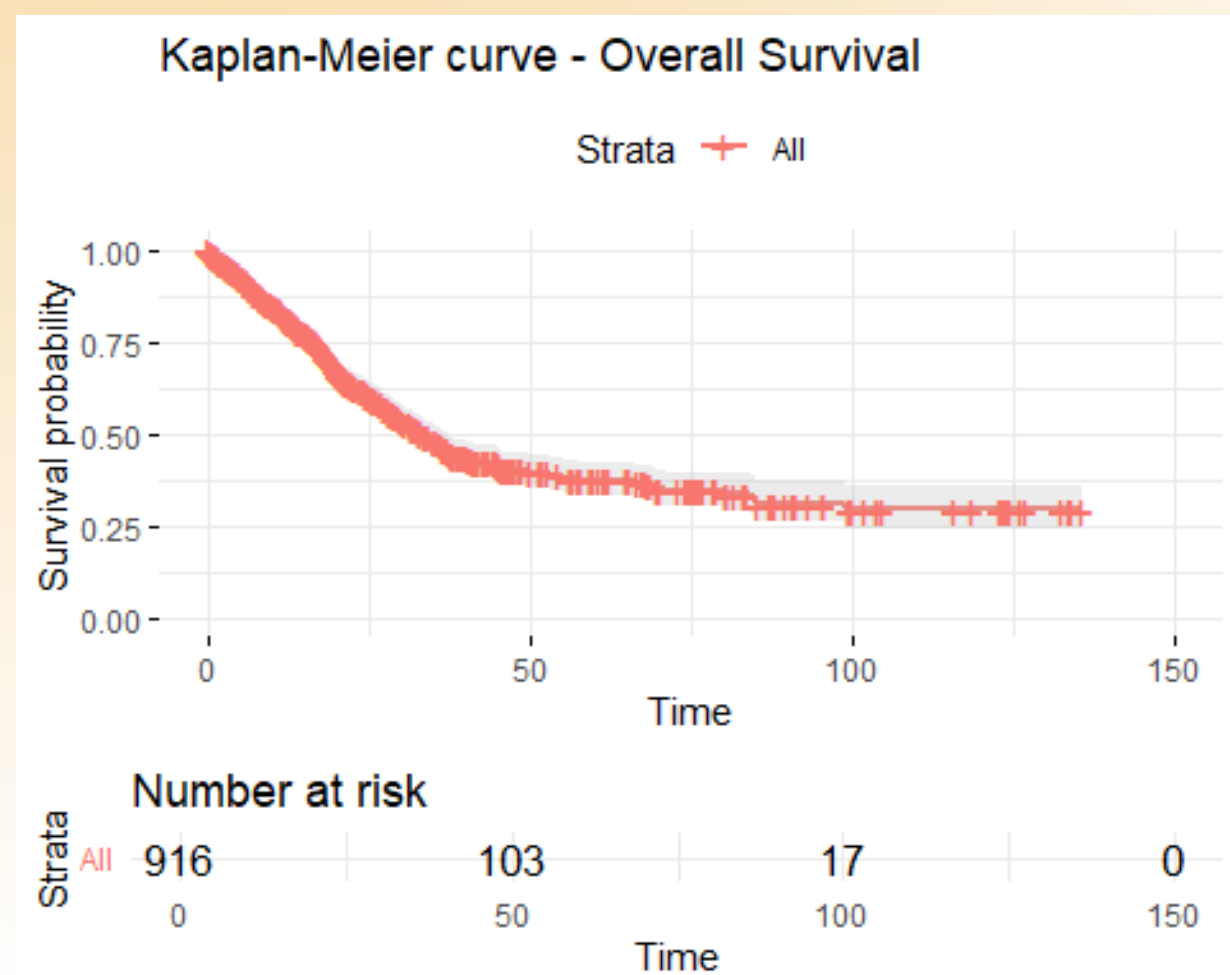# LOG RANK TEST

```
=== Log-Rank Test per ASA3 ===
Call:
survdiff(formula = Surv(IntervalOD, Death) ~ ASA3, data = data)

          N Observed Expected (O-E)^2/E (O-E)^2/V
ASA3=0 571      228      265      5.12      16.4
ASA3=1 345      158      121     11.19      16.4

 Chisq= 16.4  on 1 degrees of freedom, p= 5e-05
```

```
=== Log-Rank Test per Progress ===
Call:
survdiff(formula = Surv(IntervalOD, Death) ~ Progress, data = data)

              N Observed Expected (O-E)^2/E (O-E)^2/V
Progress=0 191       41     64.2      8.36      10.2
Progress=1 725      345    321.8      1.67      10.2

 Chisq= 10.2  on 1 degrees of freedom, p= 0.001
```

```
=== Log-Rank Test per RT ===
Call:
survdiff(formula = Surv(IntervalOD, Death) ~ RT, data = data)

        N Observed Expected (O-E)^2/E (O-E)^2/V
RT=0 818      334    337.7    0.0404     0.323
RT=1  98       52     48.3    0.2822     0.323

 Chisq= 0.3  on 1 degrees of freedom, p= 0.6
```

```
=== Log-Rank Test per NACTRT ===
Call:
survdiff(formula = Surv(IntervalOD, Death) ~ NACTRT, data = data)

            N Observed Expected (O-E)^2/E (O-E)^2/V
NACTRT=0 773      330    330.4  0.000499   0.00349
NACTRT=1 143       56     55.6  0.002963   0.00349

 Chisq= 0  on 1 degrees of freedom, p= 1
```

```
=== Log-Rank Test per TumorLOC ===
Call:
survdiff(formula = Surv(IntervalOD, Death) ~ TumorLOC, data = data)

              N Observed Expected (O-E)^2/E (O-E)^2/V
TumorLOC=0 635      263      278      0.77      2.75
TumorLOC=1 281      123      108      1.97      2.75

 Chisq= 2.7  on 1 degrees of freedom, p= 0.1
```

```
=== Log-Rank Test per AJCC_bin ===
Call:
survdiff(formula = Surv(IntervalOD, Death) ~ AJCC_bin, data = data)

              N Observed Expected (O-E)^2/E (O-E)^2/V
AJCC_bin=0 531      189      269      23.6      80.4
AJCC_bin=1 385      197      117      53.9      80.4

 Chisq= 80.4  on 1 degrees of freedom, p= <2e-16
```

COX MODEL

Global model violates the proportionality risk assumption

	chisq	df	p
Age	2.28e-01	1	0.6327
Gender	5.34e-02	1	0.8172
ASA3	4.53e+00	1	0.0333
DM	3.40e-01	1	0.5601
CAD	3.31e-01	1	0.5648
HF	8.09e-02	1	0.7761
CVA	6.89e+00	1	0.0087
CKD	4.22e-01	1	0.5162
LogCEA	9.53e-01	1	0.3290
Laparoscopic	1.12e+00	1	0.2896
TumorLOC	5.41e-01	1	0.4619
EA	9.34e-01	1	0.3338
AnesTime	3.08e-01	1	0.5788
RBC	1.16e+00	1	0.2809
Liver_Only	1.37e-04	1	0.9907
Cell_diff	3.11e+00	1	0.0779
Mucin_TYPE	8.30e-01	1	0.3624
SignetRING	1.61e+00	1	0.2052
Lymphovascularinvasion	7.59e+00	1	0.0059
perineural	1.91e+00	1	0.1665
CT	2.05e+00	1	0.1517
RT	7.88e+00	1	0.0050
NACTRT	1.57e+00	1	0.2106
Progress	7.94e+00	1	0.0048
IntervalOR	7.57e+01	1	<2e-16
AJCC_bin	5.03e-01	1	0.4782
GLOBAL	1.41e+02	26	<2e-16

COX MODEL

Hazard ratio

Variable		HR (95% CI)		p-value
Age	(N=916)	1.00 (0.98-1.01)		0.453
Gender	(N=916)	(0.76-1.18)		0.623
ASA3	(N=916)	(1.10-1.78)		0.006 **
DM	(N=916)	(0.70-1.15)		0.393
CAD	(N=916)	(0.50-1.17)		0.209
HF	(N=916)	(0.60-1.64)		0.977
CVA	(N=916)	(0.58-1.42)		0.676
CKD	(N=916)	(0.75-1.34)		0.99
LogCEA	(N=916)	(1.18-1.49)		<0.001 *
Laparoscopic	(N=916)	(0.66-0.97)		0.633
TumorLOC	(N=916)	(0.80-1.31)		0.771
EA	(N=916)	(0.65-1.16)		0.348
AnesTime	(N=916)	(1.00-1.00)		0.2
RBC	(N=916)	(1.14-1.59)		<0.001 *
Liver_Only	(N=916)	(0.98-2.93)		0.062
Cell_diff	(N=916)	(1.40-1.77)		<0.001 *
Mucin_TYPE	(N=916)	(0.59-1.39)		0.648
SignetRING	(N=916)	(0.63-1.81)		0.803
Lymphovascular invasion	(N=916)	(1.05-1.64)		0.019 *
perineural	(N=916)	(0.70-1.19)		0.529
CT	(N=916)	(0.25-0.63)		<0.001 *
RT	(N=916)	(0.81-1.49)		0.56
NACTRT	(N=916)	(0.70-1.35)		0.986
Progress	(N=916)	(0.17-0.41)		<0.001 *
IntervalOR	(N=916)	(0.82-0.92)		<0.001 *
AJCC_bin	(N=916)	(1.62 - 3.13)		<0.001 *

Events: 386; Global p-value (Log-Rank): 1.5438e-75

AIC: 4294.23; Concordance Index: 0.8
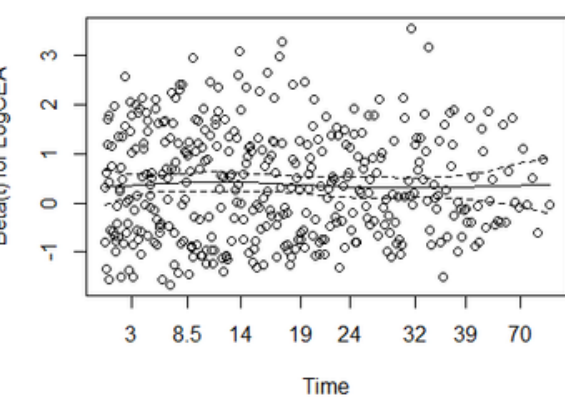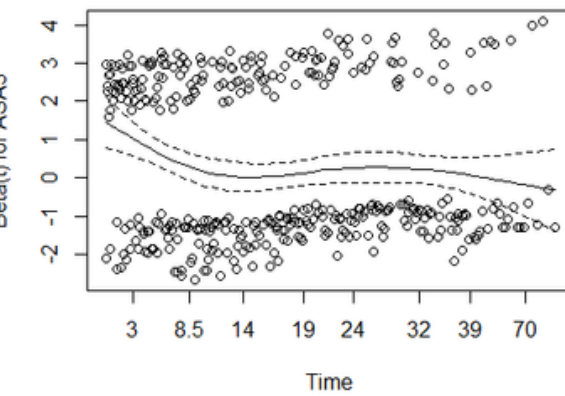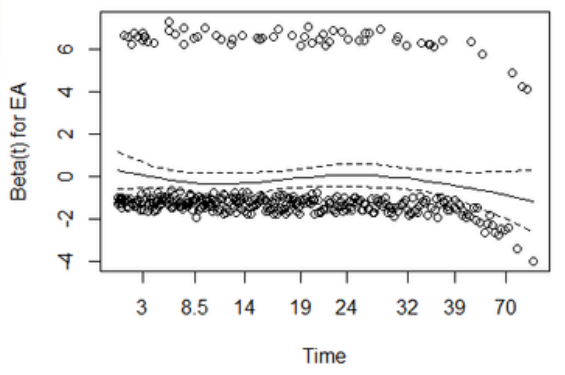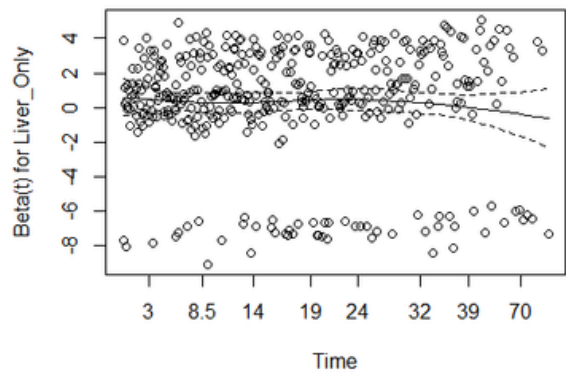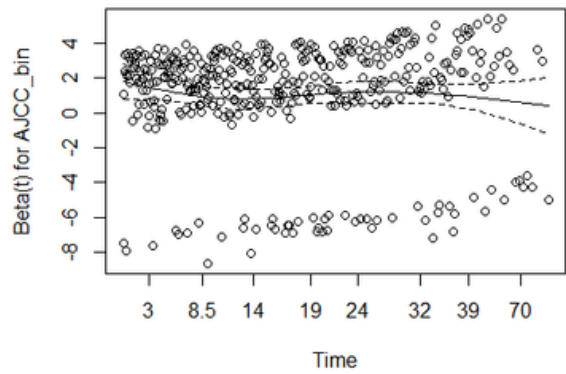
STRATIFIED COX MODEL

Statified on Progress and AJCC

	exp(coef)	exp(-coef)	lower .95	upper .95
ASA31	1.8856	0.5303	1.0363	3.431
LogCEA	1.5239	0.6562	1.1197	2.074
RBC	1.3581	0.7363	0.9312	1.981
Cell_diff	1.6863	0.5930	0.6683	4.255
Lymphovascularinvasion	0.8628	1.1590	0.4896	1.521
CT	0.4741	2.1093	0.1601	1.403

```
Concordance= 0.626  (se = 0.059 )
Likelihood ratio test= 23.16  on 6 df,      p=7e-04
Wald test            = 18.62  on 6 df,      p=0.005
Score (logrank) test = 21.61  on 6 df,      p=0.001
```

	chisq	df	p
ASA3	7.26e-04	1	0.979
LogCEA	1.04e+00	1	0.309
RBC	4.51e+00	1	0.034
Cell_diff	1.93e-01	1	0.660
Lymphovascularinvasion	5.32e+00	1	0.021
CT	3.71e+00	1	0.054
GLOBAL	1.47e+01	6	0.022

TIME DEPENDENT COX MODEL

Progress 0 used as «before»

```
                              coef  exp(coef)  se(coef)        z
ASA31                      0.27600    1.31785   0.10903    2.531
LogCEA                     0.33009    1.39110   0.05639    5.854
RBC                        0.21258    1.23686   0.07460    2.850
Cell_diff                  0.65427    1.92374   0.15529    4.213
Lymphovascularinvasion     0.27871    1.32142   0.10809    2.579
CT                        -1.49229    0.22486   0.20297   -7.352
Progress_td                2.37913   10.79554   0.19383   12.275
                          Pr(>|z|)
ASA31                      0.01136 *
LogCEA                     4.80e-09 ***
RBC                        0.00438 **
Cell_diff                  2.52e-05 ***
Lymphovascularinvasion     0.00992 **
CT                         1.95e-13 ***
Progress_td                < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                          exp(coef) exp(-coef) lower .95 upper .95
ASA31                        1.3178    0.75881    1.0643    1.6318
LogCEA                       1.3911    0.71886    1.2455    1.5537
RBC                          1.2369    0.80850    1.0686    1.4316
Cell_diff                    1.9237    0.51982    1.4189    2.6082
Lymphovascularinvasion       1.3214    0.75676    1.0691    1.6332
CT                           0.2249    4.44729    0.1511    0.3347
Progress_td                 10.7955    0.09263    7.3834   15.7845

Concordance= 0.772  (se = 0.013 )
Likelihood ratio test= 355.2  on 7 df,    p=<2e-16
Wald test            = 294.6  on 7 df,    p=<2e-16
Score (logrank) test = 330.6  on 7 df,    p=<2e-16
```

TIME DEPENDENT COX MODEL

Progress 0 used as «before»

	chisq	df	p
ASA3	6.051	1	0.0139
LogCEA	0.604	1	0.4372
RBC	0.170	1	0.6798
Cell_diff	5.245	1	0.0220
Lymphovascularinvasion	5.949	1	0.0147
CT	1.813	1	0.1782
Progress_td	5.743	1	0.0166
GLOBAL	22.943	7	0.0017

RIDGE PENALIZED COX MODEL

COEFFICIENTS: represent covariates'effects on log(risk)

ASA3: $0.243 \rightarrow HR \approx \exp(0.243) \approx 1.275$

LogCEA: $0.340 \rightarrow HR \approx \exp(0.340) \approx 1.405$

RBC: $0.211 \rightarrow HR \approx \exp(0.211) \approx 1.235$

Cell_diff: $0.707 \rightarrow HR \approx \exp(0.707) \approx 2.028$

Lymphovasc: $0.343 \rightarrow HR \approx \exp(0.343) \approx 1.409$

CT: $-1.535 \rightarrow HR \approx \exp(-1.535) \approx 0.216$

Progress_td: $2.410 \rightarrow HR \approx \exp(2.410) \approx 11.13$

SIGNIFICATIVITY TEST – ADDITIVE AALAN

COEFFICIENTS: represent covariates'effects on log(risk) THAT ARE SIGNIFICANTLY DIFFERENT FROM ZERO

ASA31: $p = 0.013$

LogCEA: $p < 0.001$

RBC: $p = 0.008$

Cell_diff: $p < 0.001$

Lymphovascularinvasion: $p = 0.001$

CT: $p = 0.003$

Progress_td: $p < 0.001$

TIME VARYING EFFECT:

LOGCEA, LYMPHOVASCULARIZATION AND PROGRESS change during time, but with a constant pace

PROGRESS is significantively variable with time

CT and CELLDIFF effects are non constant

SUPERVISED LEARNING – W/O BALANCING

Model	Accuracy	Sensitivity	Specificity	AUC	Balanced Accuracy
Logistic Regression	0.590	0.667	0.472	0.642	0.569
Random Forest	0.634	0.703	0.528	0.671	0.615
XGBoost	0.628	0.676	0.556	0.661	0.616
SVM (Radial)	0.585	0.721	0.375	0.608	0.548
KNN (k=5)	0.628	0.622	0.639	0.653	0.630
KNN (k=15)	0.607	0.622	0.583	0.655	0.603

SUPERVISED LEARNING – BALANCING

Model	Accuracy	Sensitivity	Specificity	AUC	Balanced Accuracy
Logistic Regression	0.592	0.52	0.69	0.625	0.608
Random Forest	0.62	0.549	0.75	0.65	0.65
XGBoost	0.568	0.532	0.625	0.57	0.616
SVM (Radial)	0.58	0.54	0.62	0.60	0.58
KNN (k=5)	0.585	0.586	0.583	0.605	0.585
KNN (k=15)	0.568	0.495	0.68	0.604	0.588

FEATURE IMPORTANCE

RANDOM FOREST LEAN ON:

IntervalOD	100.000000
IntervalOR	99.851676
LogCEA	91.026323
Age	82.651063
AnesTime	69.847810
RBC	25.116964
Progress	22.566055
cluster2	13.538370
Gender	12.367647
ASA3	11.487265
Lymphovascularinvasion	10.859495
DM	10.120961
TumorLOC	9.987031
RT	8.857411
perineural	8.563999
EA	8.377016

« Heuristic score, based on mean decrease in impurity or permutation impact. How much every variable contributes to reducing classification error.
IntervalOD is scale to 1, and other variables are relative to it in percentage. »

CONCLUSIONS

UNSUPERVISED LEARNING

The dataset does not exhibit natural clustering tendencies in either the original nor the projected space

SUPERVISED LEARNING AND SURVIVAL ANALYSIS

The analysis provided a set of variables that were considered important, in relation to the morbid outcome, in both the survival modeling and predictive models

CONCLUSIONS

**NEGATIVE IMPACT
DEATH-RELATED**

1. ASA3: as expected, worse categorization is related to worse outcomes
2. CEA: tumoral marker, higher values are related to worse outcomes
3. RBC: higher blood transfusion bags are more common in worse outcomes
4. Tumural cell differenciation
5. Lympho vascular invasion
6. People with 4b classification are more at risk than 4a people

**POSITIVE IMPACT
PROTECTIVE EFFECT**

1. Chemiotherapy works better than Radio therapy or NACTRT
2. Progress is strongly positive TIME RELATED BIAS
3. Interval between the operation and the re-occurrence of cancer is protective: more days are an indicator of better state

CONCLUSIONS

NON-IMPACTING VARIABLES

1. Gender: was not a predictor
2. Comorbidities: not a single comorbidity was mentioned by the different models, meaning that, in tumor-related mortality they were not a confounding or a mortality factor
3. Time under anestesia and tumor location did not have an impact
4. Age, a famous confounding variable, did not have an impact

VARIABLES THAT WOULD BE USEFUL

1. BMI : to provide more information related to patient's status
2. Time since diagnosis
3. Number of underwent operations
4. Tumor expansion in cm, in order to check actual expansion (to substitute Progress variable)

Thank's For
Watching