SUPPORT2 MEDICAL ANALYSIS

Abdullah Abdullah
Koushal Parapudi
Muhammad Ahmer

HOSPITAL INFOGRAPHICS

DATASET DESCRIPTION

PROBLEM STATEMENT

ANALYSIS TOOLS & PROCESS

RESULTS & DISCUSSION

CONCLUSION

TOOLS USED & CHALLENGES

DATASET SUMMARY & DESCRIPTION

ORIGIN

The dataset was extracted from a machine learning dataset repository: UCI Irvine

DATA SIZE

9105 instances, and 42 features with 11 categorical features and 31 numerical features

FEATURES

Frank Harrel, Department of Biostatistics at Vanderbilt University, Top 5 medical school in the United States

PROBLEM STATEMENT

LINEAR REGRESSION

1. Can predict the total hospital costs per patient.
2. Can predict the length of stay for the patients.

LOGISTIC REGRESSION

1. Can predict hospital death
2. Ordinal regression, 5-point scale, disability of patient

PAPER ASSOCIATED WITH DATASET

1. Dataset was released 6 months ago, has only a single citation
2. Could not compare results, the paper associated did not focus on the same the same features

PROBLEM STATEMENT: DESIGN OBJECTIVES

- PERFORM EXPLORATORY DATA ANALYSIS ON THE SUPPORT2 DATASET
- DEVELOP A REGRESSION MODEL FOR PREDICTING MEDICAL CHARGES
- DEVELOP A LOGISTIC REGRESSION MODEL FOR PREDICTING DEATH

ANALYTICAL TOOLS

VISUALIZ -ATION

Scatterplots, heatmaps, box plots graphs

CHI-SQUARE TEST

For dependency analysis among qualitative predictors

ONE/TWO-WAY ANOVA

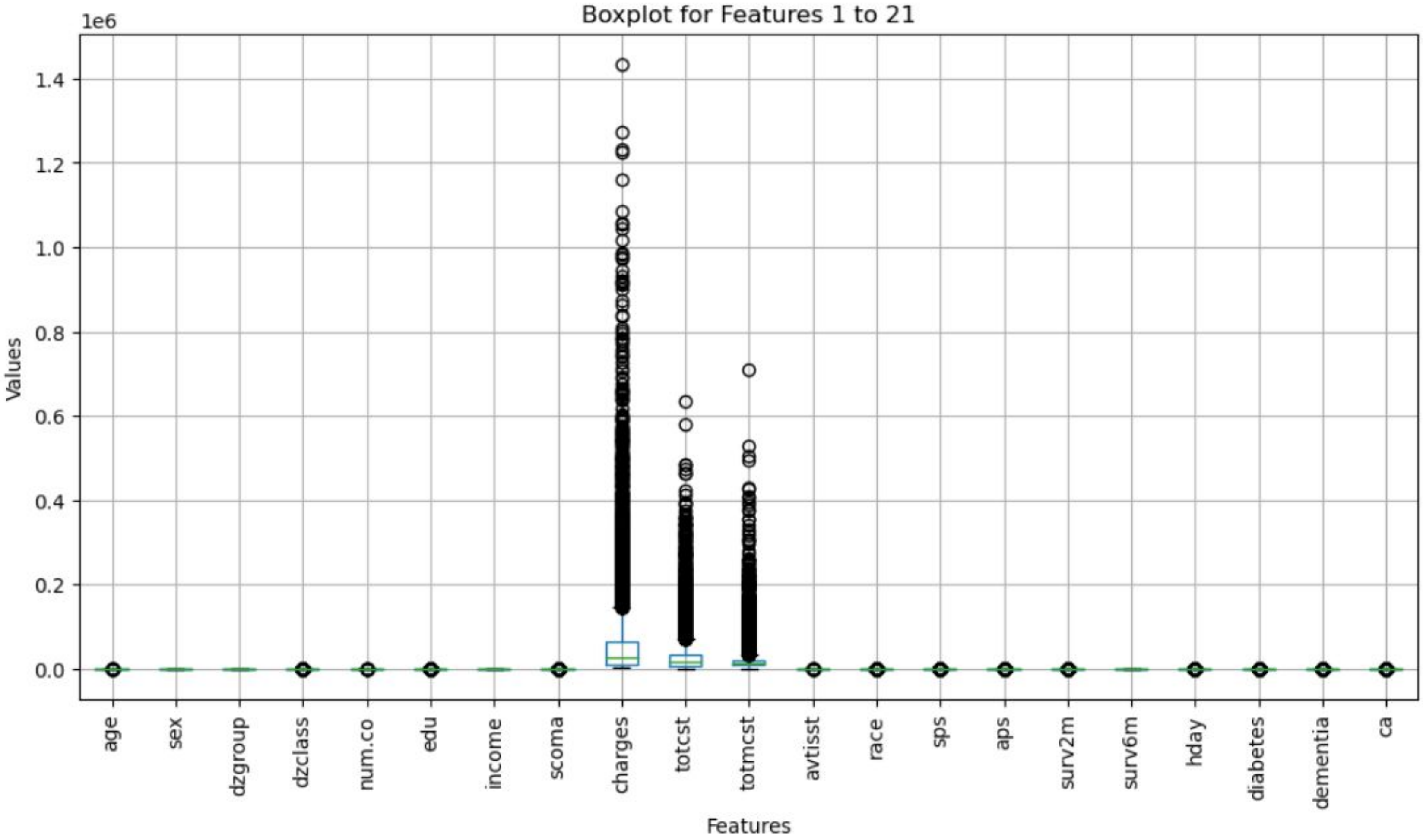
ANOVA tables for mean or factor level significance

TEST OF TWO MEANS

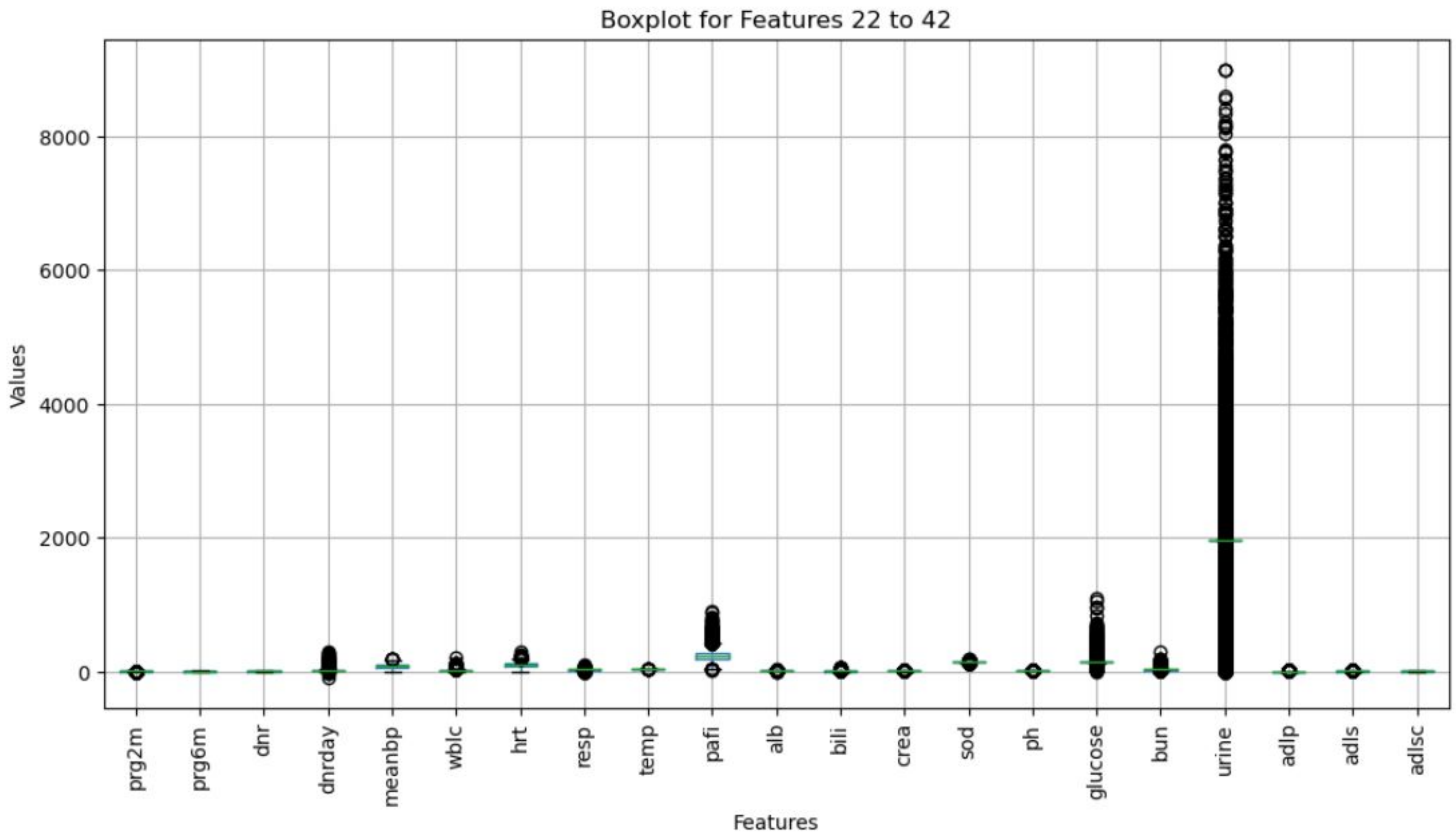
Confidence interval between test of two means

DATASET PREPROCESSING

#	Column	Non-Null Count	Dtype
---	------	--------------	-----
0	age	9105 non-null	float64
1	sex	9105 non-null	object
2	dzgroup	9105 non-null	object
3	dzclass	9105 non-null	object
4	num.co	9105 non-null	int64
5	edu	7471 non-null	float64
6	income	6123 non-null	object
7	scoma	9104 non-null	float64
8	charges	8933 non-null	float64
9	totcst	8217 non-null	float64
10	totmcst	5630 non-null	float64
11	avtisst	9023 non-null	float64
12	race	9063 non-null	object
13	sps	9104 non-null	float64
14	aps	9104 non-null	float64
15	surv2m	9104 non-null	float64
16	surv6m	9104 non-null	float64
17	hday	9105 non-null	int64
18	diabetes	9105 non-null	int64
19	dementia	9105 non-null	int64
20	ca	9105 non-null	object
21	prg2m	7456 non-null	float64
22	prg6m	7472 non-null	float64
23	dnr	9075 non-null	object
24	dnrday	9075 non-null	float64
25	meanbp	9104 non-null	float64
26	wblc	8893 non-null	float64
27	hrt	9104 non-null	float64
28	resp	9104 non-null	float64
29	temp	9104 non-null	float64
30	pafi	6780 non-null	float64
31	alb	5733 non-null	float64
32	bili	6504 non-null	float64
33	crea	9038 non-null	float64
34	sod	9104 non-null	float64
35	ph	6821 non-null	float64
36	glucose	4605 non-null	float64
37	bun	4753 non-null	float64
38	urine	4243 non-null	float64
39	adlp	3464 non-null	float64
40	adls	6238 non-null	float64
41	adlsc	9105 non-null	float64
42	death	9105 non-null	int64
43	hospdead	9105 non-null	int64
44	sfdm2	7705 non-null	object



Boxplot for Features 1 to 21
1e6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Values
age
sex
dzgroup
dzclass
num.co
edu
income
scoma
charges
totcst
totmcst
avtisst
race
sps
aps
surv2m
surv6m
hday
diabetes
dementia
ca
Features


Boxplot for Features 22 to 42
Values
8000
6000
4000
2000
0
prg2m
prg6m
dnr
dnrday
meanbp
wblc
hrt
resp
temp
pafi
alb
bili
crea
sod
ph
glucose
bun
urine
adlp
adls
adlsc
Features

CHI-SQUARE TEST OF INDEPENDENCE

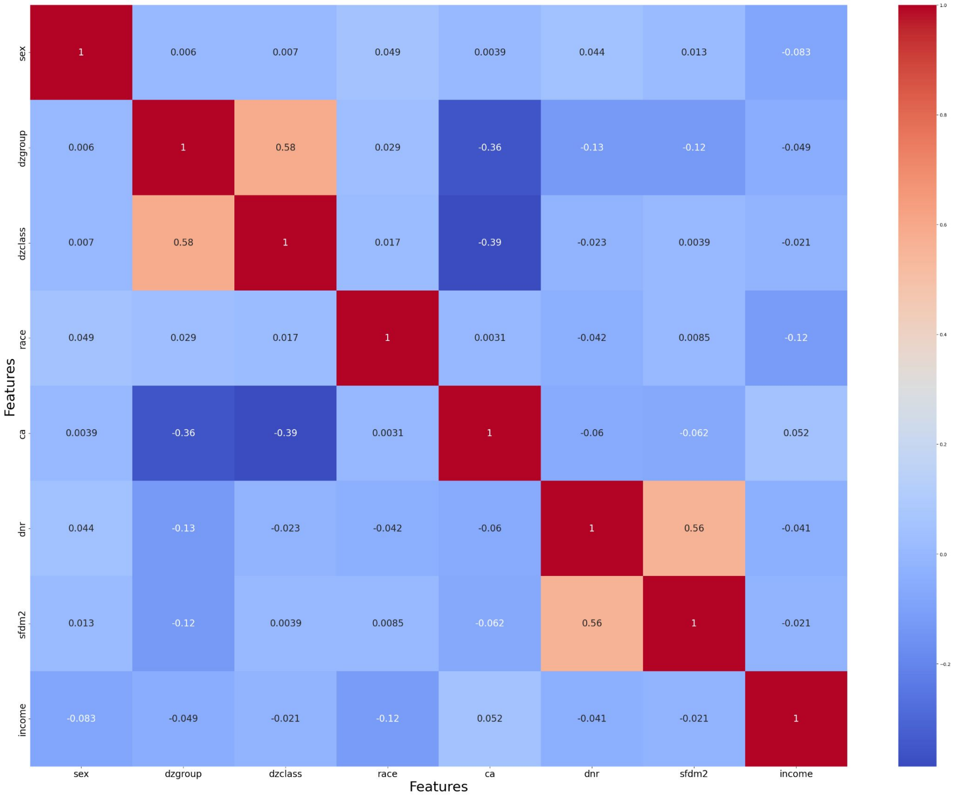
HEATMAP

Performed on the qualitative predictors

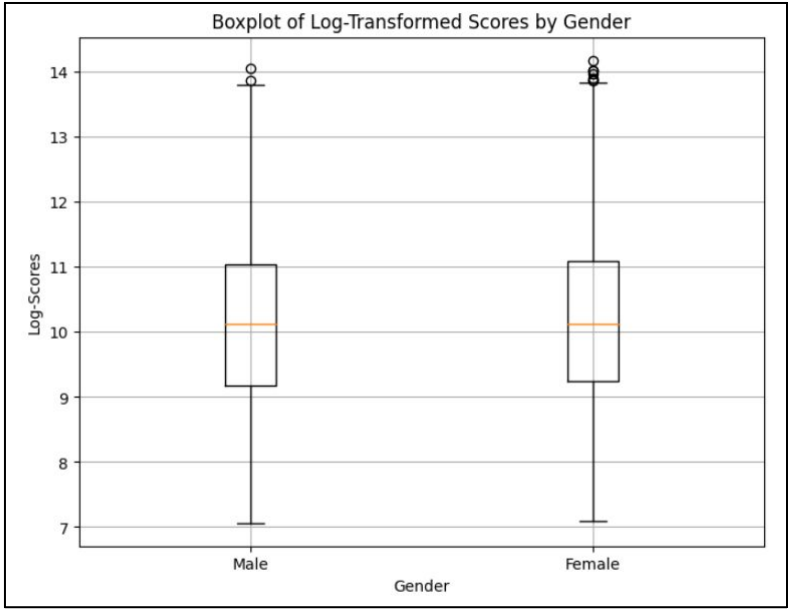
CHI-SQUARE

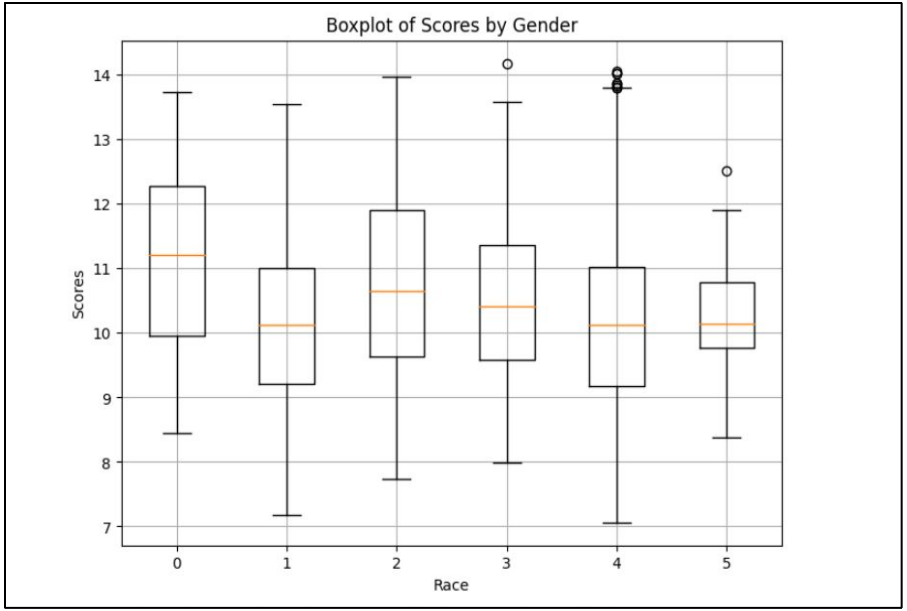
None showed sign of dependency except for ca and sex (~0.2)

Chi-square test for sex vs ca: p-value: **0.1899522053992353**



ONE-WAY ANOVA





Charges vs. Gender

pvalue=0.15282479681427144, Fail to Reject Ho

Charges vs. Race

pvalue=1.355746479339789e-24, Reject Ho

TWO-WAY ANOVA

	df	sum_sq	mean_sq	F	PR(>F)
C(ca)	2.0	3.000900e+12	1.500450e+12	149.578218	1.213537e-64
C(sex)	1.0	1.840749e+10	1.840749e+10	1.835023	1.755690e-01
C(ca):C(sex)	2.0	2.758396e+10	1.379198e+10	1.374908	2.529155e-01
Residual	9099.0	9.127394e+13	1.003121e+10	NaN	NaN

	df	sum_sq	mean_sq	F	PR(>F)
C(dzclass)	3.0	1.203796e+13	4.012654e+12	443.646463	5.096828e-269
C(sex)	1.0	8.870696e+06	8.870696e+06	0.000981	9.750173e-01
C(dzclass):C(sex)	3.0	3.126389e+09	1.042130e+09	0.115220	9.512219e-01
Residual	9097.0	8.227973e+13	9.044711e+09	NaN	NaN

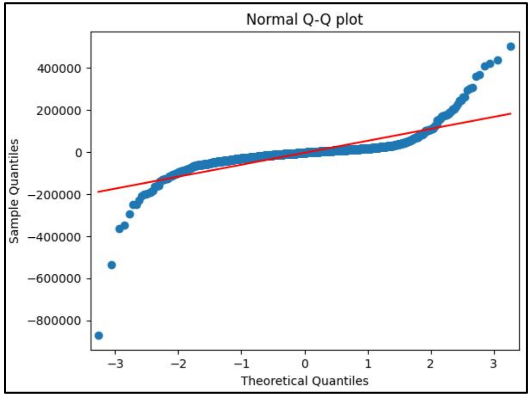
Ca vs. Sex

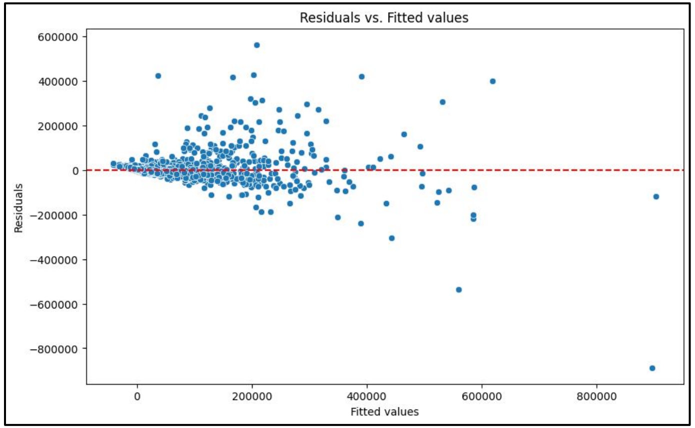
Sex vs. Dzclasses

LINEAR REGRESSION

Regression Equation:

y (charges) = 0.96 + 2246.77 * totcst + 1485.98 * hday + 683.75 * dnrday + 1174.14 * avtisst + 408.51 * bili + 1378.08 * bun + 7606.18 * edu + -6383.48 * alb + 3.36 * dnr + -419.88 * urine + -2347.03 * age + -85002.97 * dzgroup + -830.29 * surv6m + -2279.96 * sps + 286.42 * crea + -296.24 * aps + -2654.81 * wblc + 8993.67 * race + 1275.69 * prg2m + -73.21 * sfdm2 + -7601.06 * meanbp + 2221.02 * diabetes + -3872.81 * num.co + 890.42 * dzclass + -0.06 * temp + -4354.45 * totmcst + 22.46 * ph + 53908.02 * glucose + 124.08 * surv2m + -3577.79 * scoma + 2996.74 * adls + -2857.43 * adlsc + -1394.61 * sex





LINEAR REGRESSION

Regression Equation:

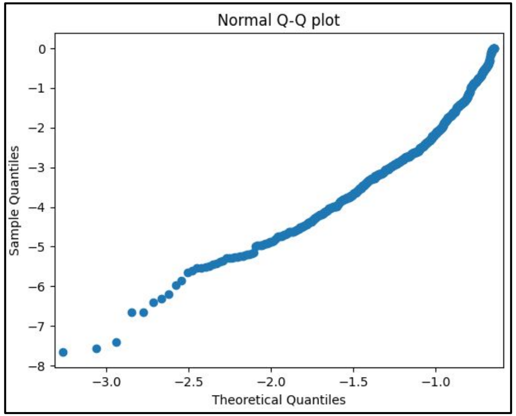
```
y (charges) = 0.96 + 2246.77 * totcst +
1485.98 * hday + 683.75 * dnrday + 1174.14 *
avtisst + 408.51 * bili + 1378.08 * bun +
7606.18 * edu + -6383.48 * alb + 3.36 * dnr +
-419.88 * urine + -2347.03 * age + -85002.97 *
dzgroup + -830.29 * surv6m + -2279.96 * sps +
286.42 * crea + -296.24 * aps + -2654.81 *
wblc + 8993.67 * race + 1275.69 * prg2m +
-73.21 * sfdm2 + -7601.06 * meanbp + 2221.02 *
diabetes + -3872.81 * num.co + 890.42 *
dzclass + -0.06 * temp + -4354.45 * totmcst +
22.46 * ph + 53908.02 * glucose + 124.08 *
surv2m + -3577.79 * scoma + 2996.74 * adls +
-2857.43 * adlsc + -1394.61 * sex
```

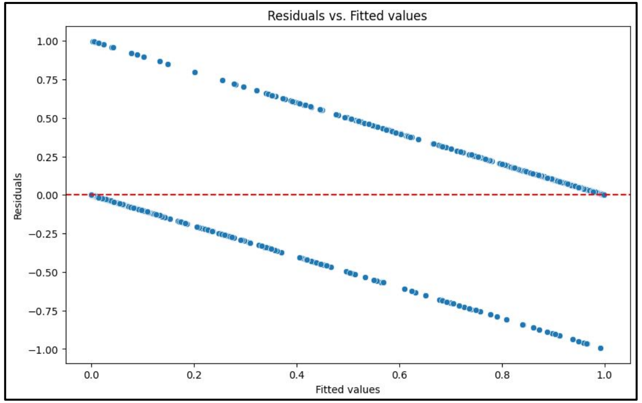
For the linear regression model, the MSE calculated was **3247121163.9039526** which is quite high and can be attributed to the noisiness of the data and the many outliers. As for R-squared adjusted, the value was **72.5%** which is decent and suggests that the linear regression can work well enough despite the high MSE.

LOGIT FUNCTION MODEL

Logistic Regression Equation:

P(Y=1) = + 0.0016 * age + -0.0912 * sex + -0.0170 * dzgroup + 0.1725 * dzclass + -0.1074 * num.co + 0.0352 * edu + 0.0133 * income + 0.0046 * scoma + 0.0000 * charges + 0.0000 * totcst + -0.0000 * totmcst + 0.0884 * avtisst + -0.1130 * race + 0.0237 * sps + 0.0198 * aps + -1.1938 * surv2m + 1.3988 * surv6m + 0.0013 * hday + -0.0289 * diabetes + -0.0470 * dementia + -0.0364 * ca + -0.9287 * prg2m + 0.9210 * prg6m + -0.7990 * dnr + 0.0178 * dnrday + 0.0036 * meanbp + -0.0035 * wblc + 0.0002 * hrt + -0.0071 * resp + -0.0014 * temp + 0.0003 * pafi + -0.0518 * alb + 0.0131 * bili + 0.0184 * crea + 0.0007 * sod + -1.0586 * ph + -0.0001 * glucose + -0.0014 * bun + -0.0000 * urine + -0.0963 * adlp + -0.1554 * adls + 0.1694 * adlsc + 14.7326 * death + -1.2396 * sfdm2

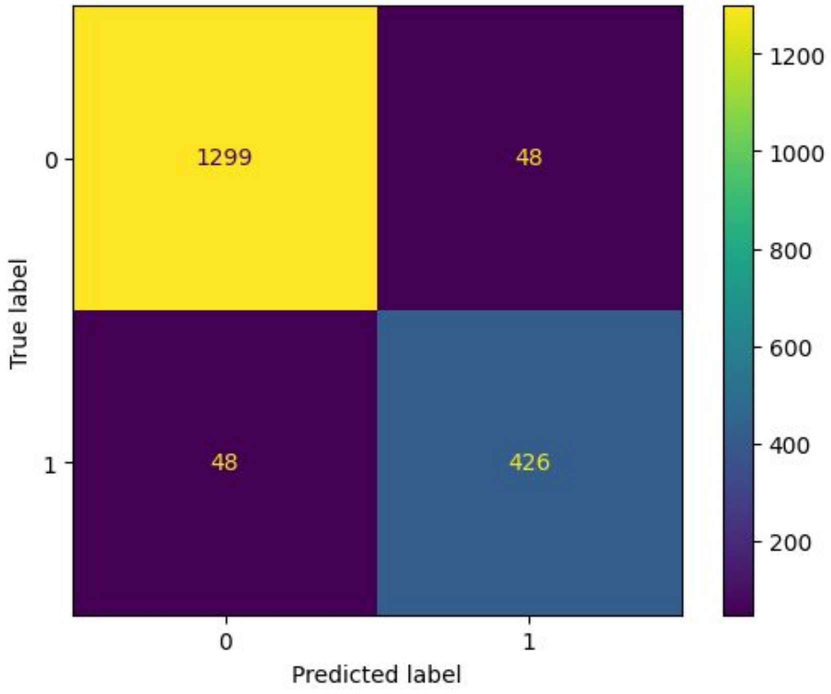




LOGIT FUNCTION MODEL

Logistic Regression Equation:

P(Y=1) = + 0.0016 * age + -0.0912 * sex + -0.0170 * dzgroup + 0.1725 * dzclass + -0.1074 * num.co + 0.0352 * edu + 0.0133 * income + 0.0046 * scoma + 0.0000 * charges + 0.0000 * totcst + -0.0000 * totmcst + 0.0884 * avtisst + -0.1130 * race + 0.0237 * sps + 0.0198 * aps + -1.1938 * surv2m + 1.3988 * surv6m + 0.0013 * hday + -0.0289 * diabetes + -0.0470 * dementia + -0.0364 * ca + -0.9287 * prg2m + 0.9210 * prg6m + -0.7990 * dnr + 0.0178 * dnrday + 0.0036 * meanbp + -0.0035 * wblc + 0.0002 * hrt + -0.0071 * resp + -0.0014 * temp + 0.0003 * pafi + -0.0518 * alb + 0.0131 * bili + 0.0184 * crea + 0.0007 * sod + -1.0586 * ph + -0.0001 * glucose + -0.0014 * bun + -0.0000 * urine + -0.0963 * adlp + -0.1554 * adls + 0.1694 * adlsc + 14.7326 * death + -1.2396 * sfdm2



For the logistic regression, we had better luck as our misclassification rate was ~0.0527 and our accuracy rate was ~0.9473 with a confusion matrix that looked like this:

DISCUSSION

Our resulting logistic regression had a decent value of R-squared adjusted of 72.5% but had a very high MSE > 3*10^9 which we concluded to be due to the data's very high number of outliers since this is data regarding critically ill patients and as such, we expect there to be many outliers. On the other hand, the logistic model predicting hospital death performed very well with an accuracy of ~94.7% on the testing partition of the dataset.

Tools used

Python

Due to versatility and availability of statistical libraries for analysis.

Pandas

Library to read and manipulate the dataset.

Numpy

For manipulation of data arrays.

SKLearn

For encoders, train test split, and analysis tools such as ANOVA.

Scipy stats

For different statistical tests such as t-test.

Statsmodels

For the regression models that were used.

CHALLENGES

A — **I**
Missing values in the data

B — **II**
Many outliers in the data

C — **III**
Large MSE for linear regression

D — **IV**
Lack of research using the same data

Conclusion

01

Sex and cancer are correlated, but sex is not significant predictor of charges.

02

Type of disease and cancer are significant predictors of charges.

03

Linear model is not suitable for predicting charges in this highly noisy dataset.

04

Logistic model for predicting hospital death is good (94.7% accuracy).

Future work

Clean data

Remove outliers and do necessary transformations.

Other models

Such as polynomial regression or machine learning models.

New tasks

Such as predicting cancer given other features, or predicting length of hospital stay.

Comparison

Compare against findings of other studies using same dataset.

THANK YOU FOR WATCHING

Abdullah Abdullah:
b00090508
Koushal Parapudi:
b00087520
Muhammad Ahmer:
b00087698