Question 1:

1.1

Import the medical cost data from the following Kaggle link https://www.kaggle.com/datasets/mirichoi0218/insurance into R.

Assignment 3.R* insurance

Filter

	age	sex	bmi	children	smoker	region	charges	isSmoker	healthyBMI	hasChildren
1	19	female	27.900	0	yes	southwest	16884.924	1	0	1
2	18	male	33.770	1	no	southeast	1725.552	0	0	0
3	28	male	33.000	3	no	southeast	4449.462	0	0	0
4	33	male	22.705	0	no	northwest	21984.471	0	1	1
5	32	male	28.880	0	no	northwest	3866.855	0	0	1
6	31	female	25.740	0	no	southeast	3756.622	0	0	1
7	46	female	33.440	1	no	southeast	8240.590	0	0	0
8	37	female	27.740	3	no	northwest	7281.506	0	0	0
9	37	male	29.830	2	no	northeast	6406.411	0	0	0
10	60	female	25.840	0	no	northwest	28923.137	0	0	1
11	25	male	26.220	0	no	northeast	2721.321	0	0	1
12	62	female	26.290	0	yes	southeast	27808.725	1	0	1
13	23	male	34.400	0	no	southwest	1826.843	0	0	1
14	56	female	39.820	0	no	southeast	11090.718	0	0	1
15	27	male	42.130	0	yes	southeast	39611.758	1	0	1
16	19	male	24.600	1	no	southwest	1837.237	0	1	0
17	52	female	30.780	1	no	northeast	10797.336	0	0	0

1.2

Multiple linear regression with Charges being the dependent variable:

$Charges_i$= -2,136.454 +266.544*age_i + 23,892.570*$isSmoker_i - 3{,}145.089 * healthyBMI_i + 1{,}037.773 * hasChildren_i + \epsilon_i$

1.3

Export the regression results to MS Word using the stargazer library.

Multiple linear regression (charges~ age + isSmoker +healthyBMI + hasChildren)	
	Dependent variable:
	charges
age	266.544***
	(12.264)
isSmoker	23,892.570***
	(425.123)
healthyBMI	-3,145.089***
	(459.968)
hasChildren	1,037.773***
	(346.847)
Constant	-2,136.454***
	(558.210)
Observations	1,338
R^2	0.733
Adjusted R^2	0.732
Residual Std. Error	6,271.992 (df = 1333)
F Statistic	912.840*** (df = 4; 1333)
Note:	*p**p***p<0.01

1.4

What is the estimated average medical costs for a 40 year old, non-smoker individual with one child and a BMI of 19?

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 19 is 6,417.984

1.5

How does your answer to the above question change if the individual had a BMI of 27?

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 27 is 9,563.073

1.6

Multiple linear regression with log(charges) being the dependent variable:

$log(charges_i)$= 7.299 +0.035*age_i + 1.546*$isSmoker_i - 0.118 * healthyBMI_i + 0.227 * hasChildren_i + \epsilon_i$

1.7

Export the regression results to MS Word using the stargazer library.

Multi linear regression with log(charges) being a dependent variable	
	Dependent variable:
	log(charges)
age	0.035^{***}
	(0.001)
isSmoker	1.546^{***}
	(0.031)
healthyBMI	-0.118^{***}
	(0.033)
hasChildren	0.227^{***}
	(0.025)
Constant	7.299^{***}
	(0.040)
Observations	1,338
R^2	0.757
Adjusted R^2	0.756
Residual Std. Error	0.454 (df = 1333)
F Statistic	$1,036.758^{***}$ (df = 4; 1333)
Note:	*p**p***p<0.01

1.8

With the new model what is the estimated average medical costs for a 40 year old, non-smoker individual with one child and a BMI of 19?

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 19 is 6,697.871

1.9

How does your answer to the above question change if the individual had a BMI of 27?

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 27 is 7,534.035

1.10

As an individual ages by 10 years by what percentage do their medical costs increase?

With age coefficient being 0.035, every 10 years of age results a change of 35% in log(charges).

In another word, there’s 41.9% in actual charges change for every 10-year increase in age.

```
[1] "The percentage difference in charges for an individual that ages by 10 years:"
> print(perc_diff)
       1
41.95096
```

1.11

Finally, estimate the following multiple regression model with the log(charges) as the dependent variable, and log(age) as one of the predictors:

log(chargesi) = β0 + β1log(agei) + β2isSmokeri + β3healthyBMIi + β4hasChildreni + εi

$log(charges_i)$= 4.16 + 1.264*$\log(age)_i$ + 1.544*$isSmoker_i - 0.122 * healthyBMI_i +$ $0.168 * hasChildren_i + \epsilon_i$

1.12

Export the regression results to MS Word using the stargazer library.

Multiple linear regression log(charges)~ log(age)+ isSmoker + healthyBMI +hasChildren

	Dependent variable:
	log(charges)
log(age)	1.264***
	(0.032)
isSmoker	1.544***
	(0.031)
healthyBMI	-0.122***
	(0.033)
hasChildren	0.168***
	(0.025)
Constant	4.160***
	(0.116)
Observations	1,338
R^2	0.756
Adjusted R^2	0.755
Residual Std. Error	0.455 (df = 1333)
F Statistic	1,033.279*** (df = 4; 1333)
Note:	*p**p***p<0.01

1.13

With the new model what is the estimated average medical costs for a 40 year old, non-smoker individual with one child and a BMI of 19?

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 19 is 7,104.192

1.14

How does your answer to the above question change if the individual had a BMI of 27?

Estimation for average medical costs for a 40-year-old non -smoker individual with one child and a BMI of 27 is 8,029.289

1.15

What does the parameter β1 represent in this model?

The coefficient $\beta 1=1.264$ indicates that a 1% increase in age is associated with approximately a 1.264% increase in charges due to the use of logarithm for both charges and age.

Question 2:

2.1

Import the bank loan default data set.

Assignment 3.R* × | d ×

Filter

	Default	Checking_amount	Term	Credit_score	Gender	Marital_status	Car_loan	Personal_loan	Home_loan	Education_loan	Emp_status	Amount	Saving_amount
1	0	988	15	796	Female	Single	1	0	0	0	employed	1536	3455
2	0	458	15	813	Female	Single	1	0	0	0	employed	947	3600
3	0	158	14	756	Female	Single	0	1	0	0	employed	1678	3093
4	1	300	25	737	Female	Single	0	0	0	1	employed	1804	2449
5	1	63	24	662	Female	Single	0	0	0	1	unemployed	1184	2867
6	0	1071	20	828	Male	Married	1	0	0	0	employed	475	3282
7	0	-192	13	856	Male	Single	1	0	0	0	employed	626	3398
8	0	172	16	763	Female	Single	1	0	0	0	employed	1224	3022
9	0	585	20	778	Female	Single	1	0	0	0	unemployed	1162	3475
10	1	189	19	649	Male	Married	1	0	0	0	employed	786	2711
11	1	214	19	742	Male	Married	0	0	0	1	employed	1270	2925
12	0	710	18	772	Female	Single	0	1	0	0	employed	1198	3694
13	0	268	18	815	Male	Married	0	1	0	0	employed	1244	3403
14	1	262	21	726	Female	Single	1	0	0	0	unemployed	977	2191
15	1	24	21	705	Female	Single	1	0	0	0	employed	1212	2965
16	0	887	18	779	Female	Single	0	1	0	0	employed	1162	3228
17	0	360	20	710	Male	Married	0	0	1	0	employed	1437	3484
18	0	571	16	779	Male	Married	1	0	0	0	employed	1318	3935
19	0	382	15	757	Male	Married	1	0	0	0	employed	910	3730
20	0	1151	13	784	Male	Single	0	1	0	0	employed	869	3315
21	0	167	19	773	Male	Married	0	1	0	0	unemployed	1200	2929
22	1	491	25	747	Male	Married	0	0	0	1	employed	868	2558

Showing 1 to 22 of 1,000 entries, 16 total columns

2.2

Calculate summary statistics for all the numerical variables in the data set.

Checking account statistical summary:

Min: -665

1st Quartile: 164.8

Median: 351.5

Mean: 362.4

3rd Quartile: 553.5

Max: 1319.0

Term statistical summary:

Min: 9

1st Quartile: 16

Median: 18

Mean: 17.82

3rd Quartile: 20

Max: 27

Credit score statistical summary:

Min: 376

1st Quartile: 725.8

Median: 770.5

Mean: 760.5

3rd Quartile: 812

Max: 1029

Amount statistical summary:

Min: 244

1st Quartile: 1016

Median: 1226

Mean: 1219

3rd Quartile: 1420

Max: 2362

Saving amount statistical summary:

Min: 2082

1st Quartile: 2951

Median: 3203

Mean: 3179

3rd Quartile: 3402

Max: 4108

Employment duration statistical summary:

Min: 0

1st Quartile: 15

Median: 41

Mean: 49.39

3rd Quartile: 85

Max: 120

Age statistical summary:

Min: 18

1st Quartile: 29

Median: 32

Mean: 31.21

3rd Quartile: 34

Max: 42

Number of credit account statistical summary:

Min: 1

1st Quartile: 1

Median: 2

Mean: 2.546

3rd Quartile: 3

Max: 9

2.3

Tabulate all the categorical variables in the data set.

Default: 300 defaulted; 700 not defaulted

Gender: 310 female; 690 male

Marital status: 548 married; 452 single

Car loan: #0: 647; #1: 353 (1 if person has a car loan, 0 otherwise)

Personal loan: #0: 526; #1: 474 (1 if person has a personal loan, 0 otherwise)

Home loan: #0: 944; #1: 56 (1 if person has a home loan, 0 otherwise)

Education loan: #0: 888; #1: 112 (1 if person has a student loan, 0 otherwise)

Employment status: 304 employed and 692 unemployed

2.4

Does there seem to be enough variation in the categorical variables to build a reliable model for loan defaults?

Yes, there seems to be enough variation in the categorical variables to build a reliable model for loan defaults.

2.5

Estimate a multiple linear regression model for loan defaults using all the variables in the data set.

Default = 3.465- 0.0003*Checking Amount + 0.014*Term – 0.001*Credit Score + 0.008*Gender (Male) – 0.042*Marital Status (Single)- 0.079*Car Loan – 0.147*Personal Loan – 0.215*Home Loan + 0.043*Education Loan + 0.052* Employment Status (Unemployed)+ 0.0001*Amount – 0.00003*Saving Amount -0.0002 * Employment Duration – 0.044*Age – 0.01*Number of credit accounts

2.6

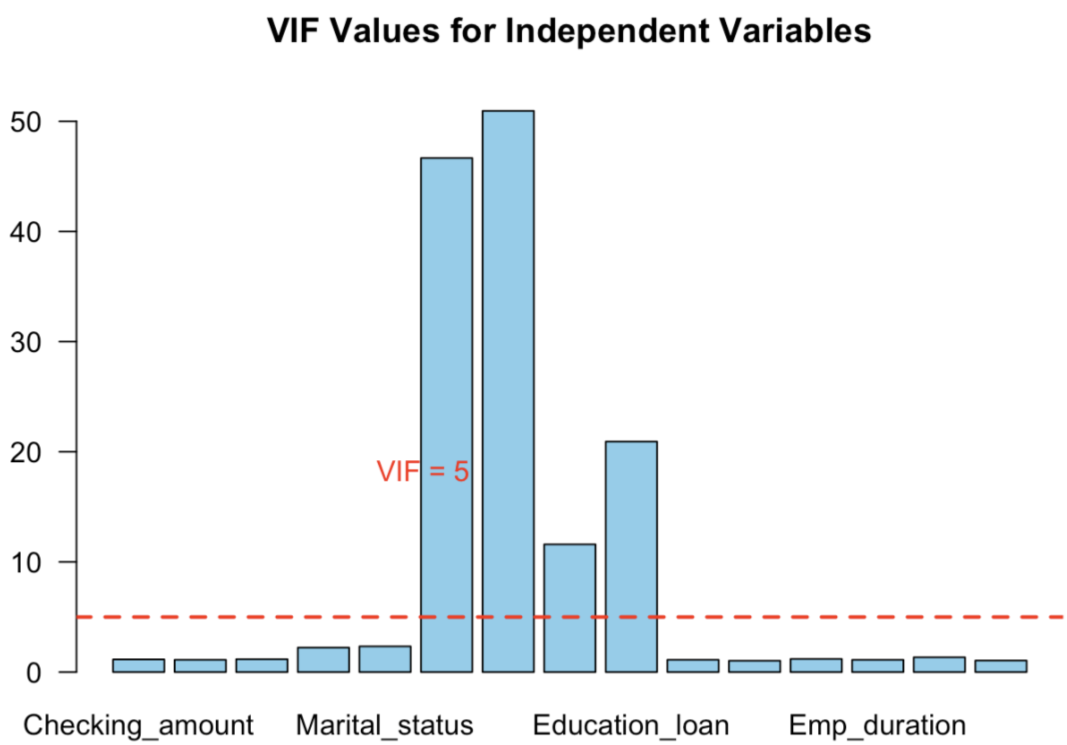
Export the regression results to MS Word using the stargazer library.

Multiple Linear Regression model for Loan Default

	Dependent variable:
	Default
Checking_amount	-0.0003***
	(0.00003)
Term	0.014***
	(0.003)
Credit_score	-0.001***
	(0.0001)
GenderMale	0.008
	(0.027)
Marital_statusSingle	0.042
	(0.026)
Car_loan	-0.079
	(0.122)
Personal_loan	-0.147
	(0.122)
Home_loan	-0.215*
	(0.126)
Education_loan	0.043
	(0.124)
Emp_statusunemployed	0.052***
	(0.020)
Amount	0.0001**
	(0.00003)
Saving_amount	-0.0003***
	(0.00003)
Emp_duration	-0.0002
	(0.0002)
Age	-0.044***
	(0.002)
No_of_credit_acc	-0.010*
	(0.005)
Constant	3.423***
	(0.186)
Observations	1,000
R^2	0.660
Adjusted R^2	0.655
Residual Std. Error	0.269 (df = 984)
F Statistic	127.320*** (df = 15; 984)
Note:	*p**p***p<0.01

2.7

Are there independent variables that exhibit a high degree of multicollinearity? Utilize the techniques that you learned in previous assignments (like correlation matrix, or VIF) to examine multicollinearity.




```
[1] "Variables with VIF > 5:"
[1] "Car_loan"       "Personal_loan"  "Home_loan"      "Education_loan"
```

By using VIF, the following independent variables exhibit a high degree of multicollinearity: Car loan, Personal loan, Home loan, Education loan

2.8

Remove all the variables that are insignificant and that are problematic due to collinearity and estimate your final multiple linear regression model with the remaining variables.

Default = 3.472 -0.0003*Checking amount + 0.016*Term -0.001*Credit score +0.028*Employment status (Unemployed) + 0.0001*Amount -0.0003*Saving account -0.047*Age

2.9

Export the regression results to MS Word using the stargazer library.

Final multiple linear regression model for Loan default

	Dependent variable:
	Default
Checking_amount	-0.0003***
	(0.00003)
Term	0.016***
	(0.003)
Credit_score	-0.001***
	(0.0001)
Emp_statusunemployed	0.028
	(0.019)
Amount	0.0001**
	(0.00003)
Saving_amount	-0.0003***
	(0.00003)
Age	-0.047***
	(0.002)
Constant	3.472***
	(0.144)
Observations	1,000
R^2	0.638
Adjusted R^2	0.635
Residual Std. Error	0.277 (df = 992)
F Statistic	249.642*** (df = 7; 992)
Note:	*p**p***p<0.01

2.10

Can the multiple linear regression model be used as a model of loan default probabilities? Illustrate the limitations of the linear regression model by making a few predictions that result in unexpected probability values.

The multiple linear regression model can't be used as a model of loan default probabilities because the dependent variable should've fallen into the valid range of probabilities, which is [0,1]. A probability greater than 1 or less than 0 is not valid. Below is the result for example test.

1	2	3	4	5
1.461205	-0.171478	-1.135992	1.067289	1.591002

Question 3:

3.1

Estimate a multiple logistic regression model for bank loan defaults, using the same variables that you used in your final multiple linear regression model in the previous question.

$\boldsymbol{log(\frac{P(Default=1)}{P(Default=0)})}$**=** 38.294 -0.005*Checking amount + 0.178*Term -0.012*Credit score +0.485*Employment status (Unemployed) +0.0005*Amount -0.0005*Saving amount - 0.626*Age

3.2

Export the regression results to MS Word using the stargazer library.

Multiple Logistic Regression for Loan Default

	Dependent variable:
	Default
Checking_amount	-0.005^{***}
	(0.001)
Term	0.178^{***}
	(0.048)
Credit_score	-0.012^{***}
	(0.002)
Emp_statusunemployed	0.485
	(0.306)
Amount	0.0005
	(0.0005)
Saving_amount	-0.005^{***}
	(0.001)
Age	-0.626^{***}
	(0.059)
Constant	38.294^{***}
	(3.616)
Observations	1,000
Log Likelihood	-168.103
Akaike Inf. Crit.	352.206
Note:	*p**p***p<0.01

3.3

Remove any insignificant variables to arrive at your final logistic regression model. Estimate your final logistic regression model to answer the questions below.

$\boldsymbol{log}\left(\frac{P(Default=1)}{P(Default=0)}\right)$**=** 38.848 -0.005*Checking amount + 0.175*Term -0.011*Credit score - 0.005*Saving amount -0.629*Age

3.4

Export the final logistic regression model results to MS Word using the stargazer library.

Final Multiple Logistic model for Loan Default	
	Dependent variable:
	Default
Checking_amount	-0.005***
	(0.001)
Term	0.175***
	(0.047)
Credit_score	-0.011***
	(0.002)
Saving_amount	-0.005***
	(0.001)
Age	-0.629***
	(0.059)
Constant	38.848***
	(3.511)
Observations	1,000
Log Likelihood	-169.985
Akaike Inf. Crit.	351.970
Note:	*p**p***p<0.01

3.5

How does the employment status of an individual impact the probability of their loan default?

With employment status being insignificant, it doesn't affect the loan default probabilities if an individual is employed or not.

3.6

What is the difference in the probability of a loan default for an individual with a 600 credit score vs. an otherwise similar individual but with an 800 credit score?

Running an example dataset for 2 similar individuals with $500 in checking, term is 22, (credit) amount is $2000, saving amount is $2200 and age is 35, the prediction returns:

For an individual with a 600-credit score, their loan default probability is 81.89% (Very likely to default the loan) while a similar individual with a 800-credit score has 31.66% loan default (<50%, not likely to default the loan). The percentage difference for probability prediction is around 50.24%

```
[1] "Probability difference:"
> print(example4_diff)
        1
50.23768
```

Question 4:

4.1

Using R, split the data set randomly into two parts: a training data set, consisting of 70% of the observations, and a testing data set, consisting of 30% of the observations.

testing_data	300 obs. of 16 variables
training_data	700 obs. of 16 variables

4.2

Estimate a logistic regression model with the training data set using the same variables from your final logistic regression model in the question above.

$log\left(\frac{P(Default=1)}{P(Default=0)}\right)$= 36.714 -0.004*Checking amount + 0.157*Term -0.012*Credit score -0.004*Saving amount -0.63*Age

4.3

Export the training set regression results to MS Word using the stargazer library.

Training Logistic Model for Loan Default	
	Dependent variable:
	Default
Checking_amount	-0.004***
	(0.001)
Term	0.157***
	(0.058)
Credit_score	-0.012***
	(0.002)
Saving_amount	-0.004***
	(0.001)
Age	-0.630***
	(0.070)
Constant	36.714***
	(3.984)
Observations	700
Log Likelihood	-121.585
Akaike Inf. Crit.	255.170
Note:	*p**p***p<0.01

4.4

Use the model you estimated to predict the probabilities of default for the individuals in the testing data set.

Preview of the probabilities when running training logistic model on testing dataset, all are within the [0,1] range:

```
           1            2            3            4            5            6            7            8            9           10           11           12           13           14 
2.880637e-02 8.256405e-04 9.512189e-01 3.605452e-02 8.473599e-01 1.379923e-03 8.940286e-01 1.537178e-03 2.120243e-01 1.490725e-02 3.719280e-01 3.241203e-04 2.458328e-02 1.414230e-03 
          15           16           17           18           19           20           21           22           23           24           25           26           27           28 
1.734761e-01 9.465866e-01 5.506784e-04 5.080209e-03 9.736357e-01 7.347651e-02 1.709776e-02 1.947222e-01 9.879691e-01 3.550718e-01 2.188391e-04 9.998312e-01 9.756326e-01 1.690038e-01 
          29           30           31           32           33           34           35           36           37           38           39           40           41           42 
9.978145e-01 1.118274e-03 8.926646e-01 2.601659e-01 9.896496e-01 9.967809e-01 2.058117e-01 8.011409e-01 4.963191e-05 1.463572e-02 9.236863e-01 9.453429e-01 5.940817e-02 8.862044e-01 
          43           44           45           46           47           48           49           50           51           52           53           54           55           56 
5.980782e-02 2.164910e-03 1.736762e-02 1.714303e-03 9.091111e-01 1.278345e-03 1.323568e-02 1.920146e-01 8.839045e-03 1.812230e-02 9.132167e-03 2.386374e-01 1.185323e-02 9.797300e-01 
          57           58           59           60           61           62           63           64           65           66           67           68           69           70 
2.393827e-01 6.165313e-03 7.816698e-01 2.087506e-02 7.192528e-04 2.996933e-01 9.199959e-01 1.601381e-02 2.452279e-01 9.816567e-01 9.125978e-01 9.943505e-01 7.432784e-02 4.949380e-04 
          71           72           73           74           75           76           77           78           79           80           81           82           83           84 
6.930184e-01 1.383530e-01 5.484040e-01 9.963305e-01 3.461864e-02 2.762494e-03 5.645399e-01 1.924228e-01 3.838959e-01 8.662519e-01 2.619195e-02 6.249905e-04 9.985093e-01 4.196991e-01 
          85           86           87           88           89           90           91           92           93           94           95           96           97           98 
3.074586e-04 2.710876e-01 1.349154e-03 1.233351e-03 1.824558e-02 9.538068e-01 1.961596e-03 7.657470e-05 4.814953e-03 1.761578e-03 1.756703e-01 4.004654e-02 4.284871e-03 1.717087e-02 
          99          100          101          102          103          104          105          106          107          108          109          110          111          112 
9.115140e-01 9.183266e-01 1.030863e-04 1.051323e-04 2.077264e-03 1.795700e-02 4.849204e-01 1.329228e-01 9.834402e-01 6.398384e-01 8.467795e-01 7.677518e-04 9.716509e-02 6.204528e-01 
         113          114          115          116          117          118          119          120          121          122          123          124          125          126 
6.457892e-01 1.806705e-02 1.919234e-02 8.577662e-01 9.482692e-01 2.295922e-03 7.598168e-02 2.565309e-03 9.998475e-01 9.218181e-01 2.028268e-02 1.454973e-03 8.383059e-03 4.885847e-02 
         127          128          129          130          131          132          133          134          135          136          137          138          139          140 
7.841713e-03 9.973980e-01 8.817654e-01 9.992270e-01 1.128668e-01 7.746323e-02 9.941714e-01 5.458799e-01 7.193386e-05 8.300901e-01 9.615248e-01 6.892484e-02 9.905661e-01 2.896512e-03 
         141          142          143          144          145          146          147          148          149          150          151          152          153          154 
2.547878e-03 1.261588e-02 2.420282e-01 2.091331e-04 6.050190e-01 1.014460e-02 4.475424e-02 1.510533e-02 4.768276e-03 3.693801e-02 5.605390e-01 9.955394e-01 9.330187e-03 2.977473e-02 
         155          156          157          158          159          160          161          162          163          164          165          166          167          168 
1.852936e-02 9.995526e-01 2.089265e-01 6.592358e-02 7.203815e-01 1.316913e-04 9.920840e-01 8.672370e-03 7.985511e-01 1.587542e-02 8.719060e-05 6.505573e-02 5.114730e-04 9.831176e-01 
```

4.5

Assuming a cutoff probability at 70% (i.e. if the predicted probability is greater than or equal to 70% the loan will be considered as default) create a classification table for your model.

Classification table:

```
[1] "Confusion Matrix:"
> print(confusion_matrix)
   pred_class
      0   1
  0 196   3
  1  23  78
```

4.6

Using the classification table above, calculate the accuracy rate of your model as the ratio of correctly predicted outcomes over the total possible outcomes.

The accuracy rate of my model as the ratio of correctly predicted outcomes over the total possible outcomes is 91.33%

```
[1] "Accuracy percentage rate:"
> print(accuracy)
[1] 91.33333
```