Canadian General Social Survey analysis and modeling

Problem Set #2

Yangyang Liu (1003986984) Xiaoxi Bai (1004144749)
Mengyu Lei (1004735405) Yuika Cho (1003213186)

October 19, 2020

Abstract

There are four people working on problem set 2. In this problem set, readers can find data analysis and regression modeling according to 2017 Canadian General Social Survey (GSS in the rest of text) data set. 2017 Cycle 31 GSS data set is about family survey, and it was downloaded from University of Toronto Library and cleaned by gss_cleaning.R module provided in the problem set 2 package. Readers are able to find social/family issues from important statistical analysis below. Another goal of this problem set is to build up linear regression model in order to investigate linear relationship between Age and some other predictors.

1. Introduction

Family plays an important role in people's live. In the recent decades, Canadian keep focusing on family issues and observe that families are becoming more divergent. Hence, the goal of this survey is to do some data analysis on GSS family data as well as monitoring relationship between **Age** and other family factors by using linear model.

Before starting data analytic and linear regression modeling, we will present overview of this group work in the introduction section. In this section, readers can have a preview of structure of article from data and modeling perspective.
First of all, 2017 Canadian GSS on family data set is downloaded from University of Toronto library CHASS. Data set contains responses of sample survey with cross-sectional design from February 2017 to November 2017. The survey's target population is all person greater than 15 years old living in 10 different provinces in Canada. Targets of the survey are knowing more about Canadian families and capture potential social/family issues in the future. In raw GSS data set, there are 20602 observations and 461 different attributes (either continuous or categorical). After proceeding GSS cleaning module, it generates a tidier GSS data includes 81 different variables and 20602 records. In this article, all data analysis and linear regression modeling will be based on this tidy GSS data.
Secondly, in this article, linear regression model will be introduced and used to monitoring

relationship between **Age** and other family factors, because *Age* is a continuous variable. In addition to *Age*, there are 80 attributes left in the GSS data set, but selected predictors are *marital status*, *average working hours*, *income family*, and *self rated health.* These four predictors are carefully selected because they are four of the most important factors to a family, and they are representative for investigating family issues.
Thirdly, we will present to readers some quantitative numbers, graphs and summary of data analysis and linear regression modeling results.
Finally, conclusions will be made from data analysis and linear regression modeling results. At the same time, weaknesses and corresponding next step plan will also be presented and explained.

2. Data

2017 Canadian GSS on family data set is downloaded from University of Toronto library CHASS. The reason we select it is because this is an official survey data set by Statistic Canada, and it can reflect real Canadian family/social issues and worth to analyze. According to 2017 Canadian GSS on family data set user guide (*Reference*), there are few purposes of Canadian General Social Survey on family:

- To have a better understanding of families in Canada by answering few family matter questions. For example: How many families are there in Canada? What are their characteristics and socio-economic conditions? and etc.

- To reveal some potential social/family issues of matter of current interest.

- Predicting and analyzing social/family changes in the living conditions and well-being of Canadians over time.

Moreover, there are few **features** of GSS data set that we are going to analyze:

- The survey's target population is all person greater than 15 years old living in 10 different provinces in Canada.

- Survey is conducted through telephone, hence the population frame should be all person greater than 15 years old living in 10 different provinces in Canada and have telephone (landline and cellular).

- In the tidy GSS data set, most of attributes are categorical variables, for example, *marital status*, *average working hours*, *income family*, and *self rated health.*

At the same time, there are still existing some **drawbacks** in GSS data set:

- Since responses of survey were collected through telephone (landline and cellular), which is not a popular way nowadays. GSS Data is subject to both sampling and non-sampling errors.

- One of data features says that most of attributes are categorical, however, some continuous attributes are also be classified into different categories , for example: *income family*, *income respondent* and etc. Such kind of classification may cause some biases when doing data analysis.

- There are lots of missing values, so that we are choosing to ignore in this article.

3. Model

The goal of this article is to investigate relationship between and *Age* and *marital status*, *average working hours*, *income family*, and *self rated health*. In this part, Linear Regression Model will be briefly introduced with model definition and model diagnostics.

3.1 Model Definition

Linear regression model is a linear model that measures linear relationship between dependent variable (Y) and independent variables (X). Here Y need to be a continuous variable and X can be either continuous or categorical variable. Assume that there a n independent variables, mathematically, linear regression can be expressed as:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \epsilon \tag{1}$$

where β_i are estimated coefficient of each independent variable, interpretation of β_i: With one unit change of x_i causes β_i unit changes in y.

ϵ is residual term of this linear regression model.

3.2 Model Diagnostic

Linear regression model doesn't work all the time, there are few assumptions need to be met for linear regression model:

- Linearity: The relationship between independent variables and the mean of dependent variable is linear. This assumption can be diagnosed by checking scatter (pair) plot between x and y.

- Multicollinearity: Predictors cannot have many correlations (multicollinearity).
- Residual must be normally distributed: This can be tested by Quantile-Quantile Plot.

- Homoscedasticity: Residual must consistent across all variables, this can be tested by checking residual plot of linear regression.

- Outliers and Influential points: Check outliers and influential points that have bad impact on linear regression model.

3.3 Model Special Cases and caveats

For the linear regression model that will be used to analyze GSS data set, all independent variables (x) are categorical variables, so we hypothesize the relationship between independent variables and the mean of dependent variable is linear and there is no multicollinearity exists. At the same time, there are few caveats about the linear regression model in this article. Our model may not meet all model criterion, for example, Homoscedasticity doesn't meet or there are some bad outliers and Influential points can impact our analysis.

4. Results

In this section, data analysis and modeling results will be presented to reader. The section will be divided into two sub-sections, one is about data summary by showing box plots and some important statistic numbers (five number summary, mean and variance). The other section is about linear regression model results and summary.

4.1 Data Analysis

Figure 1

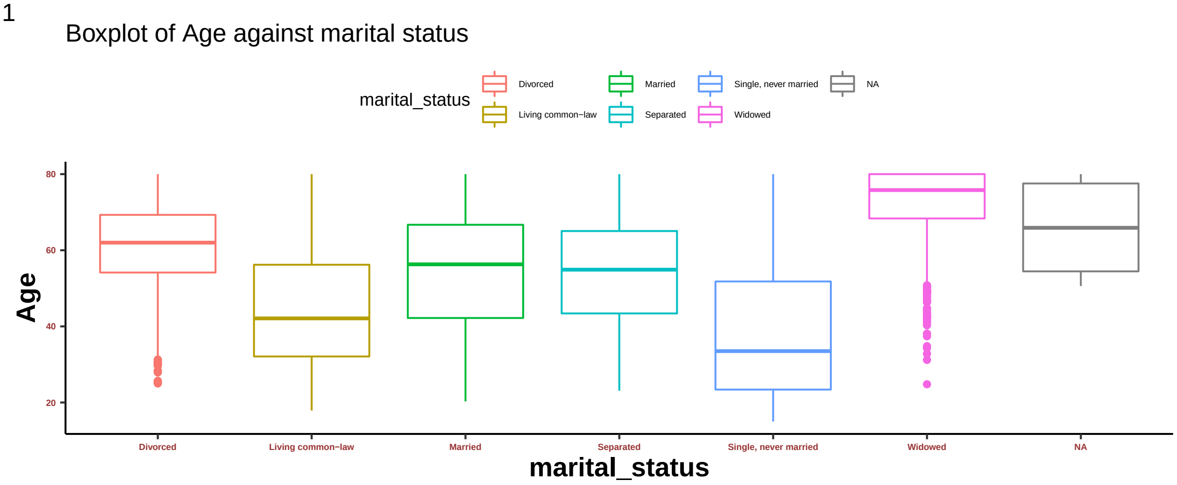


Table 1

This table shows minimum, 25% quantile, median, mean75% quantile, maximum and variance of age in each marital status's category

marital_status	Count	Minimum	Q1	Median	Mean	Q3	Maximum	VAR
Married	9501	20.3	42.20	56.3	54.90	66.70	80	219.24
Single, never married	4710	15.0	23.40	33.5	38.06	51.80	80	295.64
Living common−law	2075	17.9	32.10	42.1	44.56	56.20	80	208.99
Widowed	1899	24.8	68.35	75.8	72.99	80.00	80	71.73
Divorced	1767	25.0	54.15	62.0	61.01	69.30	80	129.58
Separated	643	23.1	43.40	54.9	54.47	65.05	80	187.11
NA	7	50.6	54.45	65.9	65.79	77.55	80	165.47

From figure 1 and table 1, we present box plot between age and marital status and statistical numbers of each marital status's category. 9500 people are married with minimum 20 years old to maximum 80 years old. There are also seven several missing values. *Widowed* have the biggest median and mean but smallest variance, from figure1, we can tell that most of widowed person are between 70 year old and 80 years old.

Figure 2

Boxplot of Age against average working hours

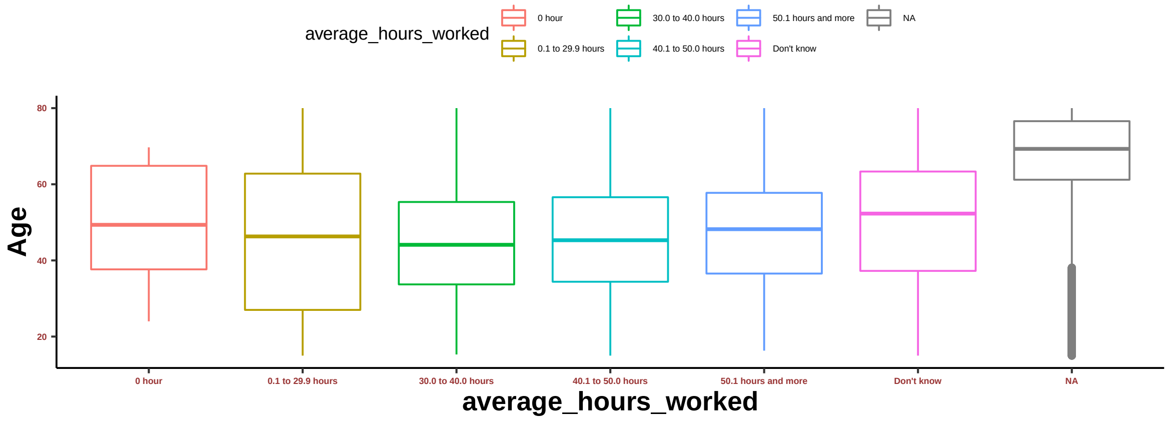
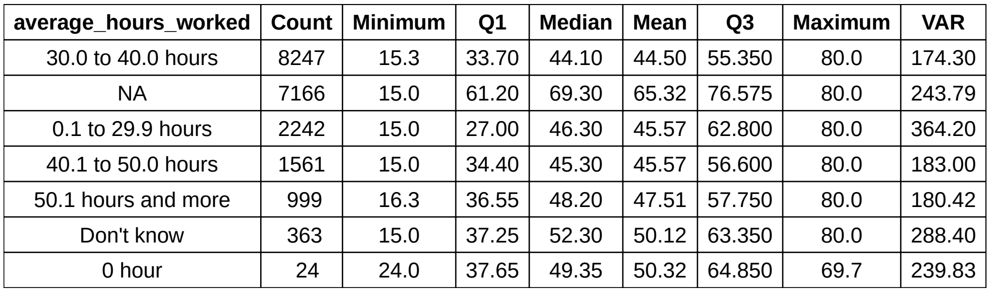


Table 2

This table shows minimum, 25% quantile, median, mean, 75% quantilemaximum and variance of age in each average working hours's category



average_hours_worked	Count	Minimum	Q1	Median	Mean	Q3	Maximum	VAR
30.0 to 40.0 hours	8247	15.3	33.70	44.10	44.50	55.350	80.0	174.30
NA	7166	15.0	61.20	69.30	65.32	76.575	80.0	243.79
0.1 to 29.9 hours	2242	15.0	27.00	46.30	45.57	62.800	80.0	364.20
40.1 to 50.0 hours	1561	15.0	34.40	45.30	45.57	56.600	80.0	183.00
50.1 hours and more	999	16.3	36.55	48.20	47.51	57.750	80.0	180.42
Don't know	363	15.0	37.25	52.30	50.12	63.350	80.0	288.40
0 hour	24	24.0	37.65	49.35	50.32	64.850	69.7	239.83

From figure 2 and table 2, we present box plot between age and average working hours and statistical numbers of each average working hours' category. 8200 people are working between 30 to 40 hours with minimum 15 years old to maximum 80 years old. There are 7100 missing values under this attributes, *average working hours* under these missing values have the biggest median and mean and variance.

Figure 3

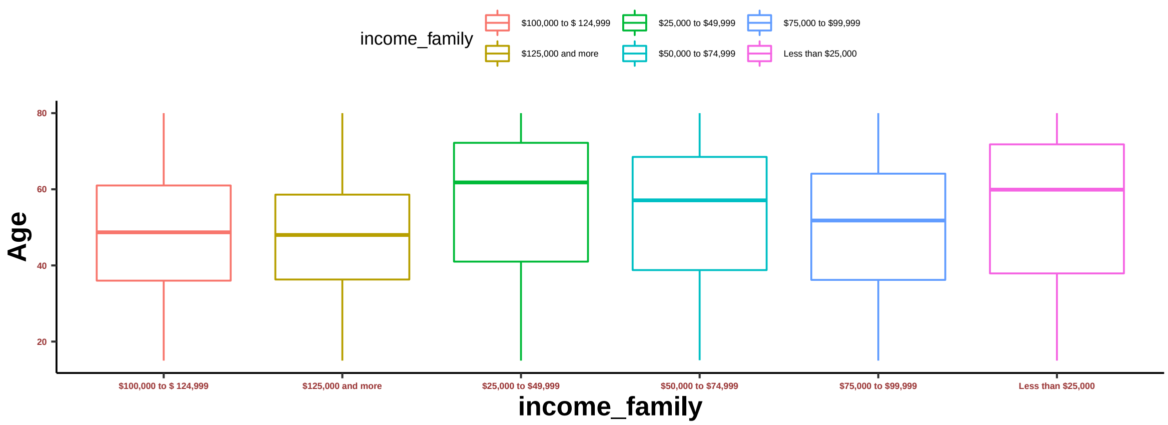

Boxplot of Age against family income

Table 3

This table shows minimum, 25% quantile, median, mean, 75% quantilemaximum and variance of age in each family income's category

income_family	Count	Minimum	Q1	Median	Mean	Q3	Maximum	VAR
$125,000 and more	4707	15.0	36.300	48.0	47.16	58.6	80	237.61
$25,000 to $49,999	4345	15.0	41.000	61.8	57.06	72.2	80	339.99
$50,000 to $74,999	3696	15.1	38.775	57.1	54.03	68.5	80	306.96
$75,000 to $99,999	2921	15.0	36.200	51.8	50.52	64.1	80	280.20
Less than $25,000	2775	15.0	37.900	59.9	55.17	71.8	80	397.70
$100,000 to $ 124,999	2158	15.0	36.000	48.7	48.65	61.0	80	251.77

From figure 3 and table 3, we present box plot between age and family income and statistical numbers of each average working hours' category. 4700 people have family income greater than $125,000 with minimum 15 years old to maximum 80 years old. There are no missing values under this attributes, *family income* among different categories has similar median, mean and variance. And from boxplot, we can tell under *family income*, age are approximately normal distributed.

Figure 4

Boxplot of Age against self rated health

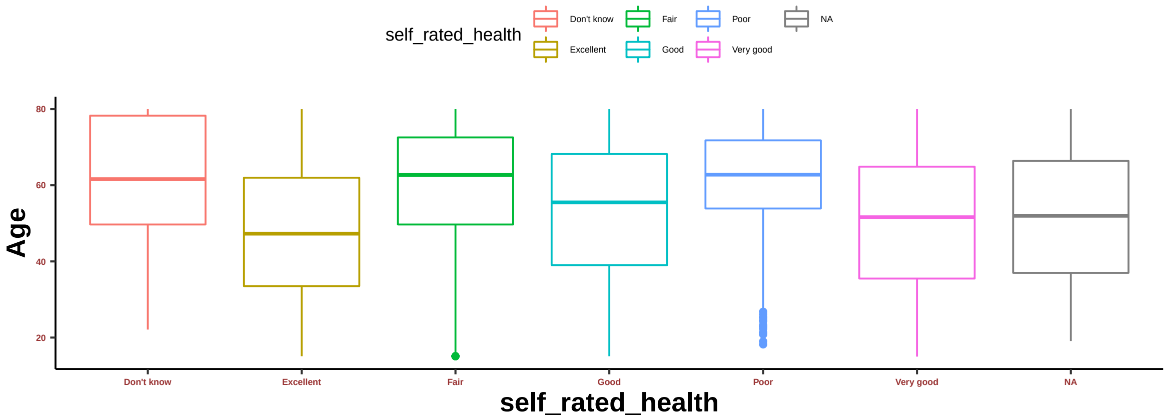


Table 4

This table shows minimum, 25% quantile, median, mean, 75% quantilemaximum and variance of age in each self rated health's category

self_rated_health	Count	Minimum	Q1	Median	Mean	Q3	Maximum	VAR
Very good	7014	15.0	35.5	51.6	50.35	64.900	80	314.05
Good	6162	15.1	39.0	55.5	53.61	68.200	80	309.98
Excellent	4376	15.1	33.5	47.3	47.87	62.000	80	298.99
Fair	2078	15.1	49.7	62.7	59.47	72.575	80	270.27
Poor	816	18.2	53.9	62.8	61.38	71.800	80	196.34
NA	99	19.1	37.0	52.0	52.35	66.400	80	293.12
Don't know	57	22.1	49.7	61.6	60.85	78.300	80	294.33

From figure 4 and table 4, we present box plot between age and self rated health and statistical numbers of each self rated health's category. 7000 people have very good health with minimum 15 years old to maximum 80 years old. There are 99 missing values under this attributes, *good* health condition has higher median, mean. And from boxplot, we can tell under most of *family income* categories, age are right-skewed.

4.2 Model Results

4.2.1 Linear Regression Coefficient

	.
(Intercept)	63.5092575
marital_statusLiving common-law	-12.8617272
marital_statusMarried	-5.4865427
marital_statusSeparated	-5.4242984
marital_statusSingle, never married	-20.0828625
marital_statusWidowed	6.2450639
average_hours_worked0.1 to 29.9 hours	-2.6885103
average_hours_worked30.0 to 40.0 hours	-4.4614063
average_hours_worked40.1 to 50.0 hours	-3.5898002
average_hours_worked50.1 hours and more	-2.2027080
average_hours_workedDon't know	0.2630130
income_family$125,000 and more	-0.0079201
income_family$25,000 to $49,999	2.4824641
income_family$50,000 to $74,999	1.8022902
income_family$75,000 to $99,999	0.5080232
income_familyLess than $25,000	0.1401072
self_rated_healthExcellent	-6.9763376
self_rated_healthFair	-2.7838257
self_rated_healthGood	-4.7882269
self_rated_healthPoor	-1.4842930
self_rated_healthVery good	-6.3120988

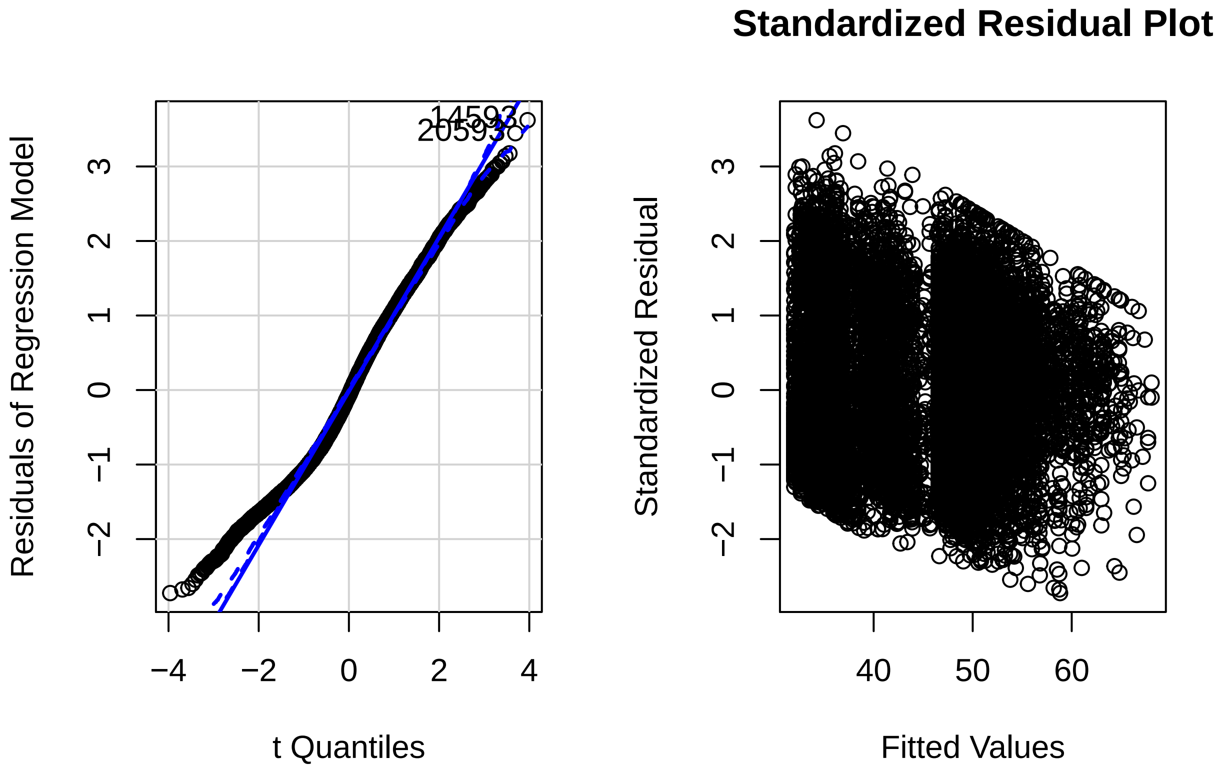
From Linear Regression Model coefficient table above, our linear regression model with estimated parameters is:

$$Age = 63.5 - 12.86 Marital_{Common\_law} + ... - 6.3 self\_rated\_health_{very\_good} \tag{2}$$

That means, for people under common law marital status, these people are about 12.86 younger than others. For people who married, they are about 5 younger than others. For people who think they have very good health condition, they are about 6 younger than others. Other predictors can be explained as the same way.

4.2.2 Linear Regression Model Diagnostic

```
## [1] 14593 20593
```



From above model Diagnostic, readers can find residuals are approximately normal distribution. However, residuals are not constant over the fitted values(*Age*), that means Homoscedasticity doesn't meet in the linear regression model.

5. Discussion

From data analysis and linear regression model results, we can conclude that *marital status* is the most important factor that impact response variable : *age*. Further, we can also conclude that for young people (less than 30 years old), many of them never married, working 0.1 to 29.9 hours and excellent health condition. For old people (more than 60 years old), many of them are widowed, doesn't have enough working hours information, and poor health condition. Whole article explained our original goal about investigating relationship between *Age* and *marital status*, *average working hours*, *income family*, and *self rated health*. In the following sub-sections, we will present some Weaknesses of our study and corresponding further steps.

5.1 Weaknesses

There are few Weaknesses that can cause our analysis less accurate or bias from real situations.

1. There are many missing data in the GSS data set as well as in our interested variables, but data imputation was not processed for those missing data.
2. Back to the survey, we know it subject to both sampling errors and non-sampling errors from GSS user document. As a result, data analysis and regression modeling also have some errors because of the survey.
3. After model diagnostics, we found that not all assumptions of linear regression model are satisfied. Also some other model assumptions need to be checked and clarified more clearly.

5.2 Next Steps

After realizing some weaknesses of our analysis in this article, we also listed further steps about how to improve performance of data analysis and regression modeling.

1. For many missing value in GSS data set, data imputation is necessary by either collecting more response or classify these missing values into existing categories.
2. In order to minimize both sampling errors and non-sampling errors according to GSS user document, it is better to diversify ways of conducting survey or investigate real biased term and present to readers.
3. For linear regression model, we will do a further diagnostics from model assumptions, model prediction (Accuracy) and bad influential points perspectives. We diagnose linear regression model has heteroscedastic (non-constant error) and non-linearity issues , we will do some model transformation in order to meet all model requirements.

Reference

Wu, Changbao. Thompson. 2020. “Sampling Theory and Practice. Springer International Publishing”.

Statistic Canada. April, 2019. “Public Use Microdata File Documentation and User’s Guide”.

Alvin C.Rencher and G. Bruce Schaalje, 2007. “Linear Models In Statistics”.

Marina Soley-Bori, 2013. “Dealing with missing data: Key assumptions and methods for applied analysis”

Rohan Alexander and Sam Caetano. 2020. “gss_cleaning.R”.