the evolution of education for 4- to 5-year-old children

produce a Data Perspective on the evolution of education for 4- to 5-year-old children. Particularly interesting for the Perspective is understanding how educational performance evolves month by month at these critical ages, considering both general education and specific subjects (e.g., literature and math, physical education). Methods should be sound and well-argued if necessary.

Load packages and import dataset

```
library(tidyverse)
library(lubridate)
library(kableExtra)



df = read.csv("01_rawdata/Zimbabwe_children_under5_interview.csv")

df$age_in_months = interval(df$child_birthday ,df$interview_date)  %/%
months(1)

unique(df$age_in_months)

##  [1] 37 40 52 47 45 46 44 49 59 43 51 57 55 38 42 54 48 56 58 36 50 41 53
39 NA

df$age_in_months_cat [df$age_in_months < 41] = 'less than 40 months'
df$age_in_months_cat [(df$age_in_months >= 41) & (df$age_in_months < 46)] =
'between 41 to 45 months'
df$age_in_months_cat [(df$age_in_months >= 46) & (df$age_in_months < 51)] =
'between 46 to 50 months'
df$age_in_months_cat [(df$age_in_months >= 51) & (df$age_in_months < 56)] =
'between 51 to 55 months'
df$age_in_months_cat [(df$age_in_months >= 56) ] = 'above 56'

df$age_in_months_cat = factor(df$age_in_months_cat , ordered =
is.ordered(df$age_in_months_cat) )

levels(df$age_in_months_cat) = c("less than 40 months" , "between 41 to 45
months" ,  "between 46 to 50 months" , "between 51 to 55 months" , "above
56")

kable(table(df$age_in_months_cat) , format = "markdown")
```

Var1	Freq
less than 40 months	429
between 41 to 45 months	518
between 46 to 50 months	500

Var1	Freq
between 51 to 55 months	536
above 56	508

Evolution of education in Literacy + Math subjects

```
# EC6: "Can (name) identify or name at least ten letters of the alphabet?"
"Yes=1/No=2/DK=8"
# EC7: "Can (name) read at least four simple, popular words?"
"Yes=1/No=2/DK=8"
# EC8: "Does (name) know the name and recognize the symbol of all numbers
from 1 to 10?" "Yes=1/No=2/DK=8"
```

```
tab1 = df %>% group_by( age_in_months_cat , EC6) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))
```

```
## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.
```

```
tab1 = tab1 [(tab1$EC6 == 1  | tab1$EC6 ==2),]
tab1$EC6 [tab1$EC6 ==1] = "Yes"
tab1$EC6 [tab1$EC6 ==2] = "No"
tab1 = tab1[!is.na(tab1$age_in_months_cat),]
kable(tab1 , format = "markdown")
```

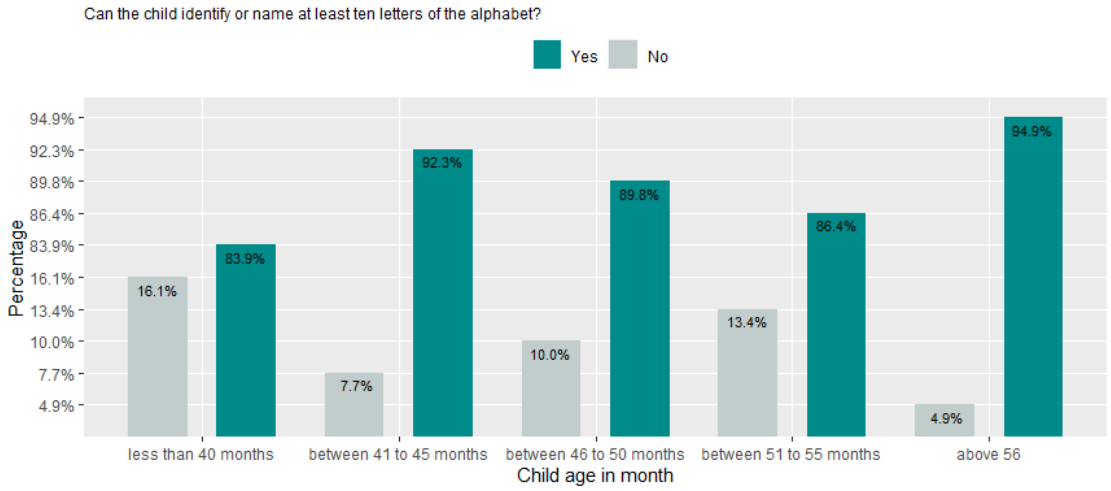
age_in_months_cat	EC6	n	perc
less than 40 months	Yes	69	16.1%
less than 40 months	No	360	83.9%
between 41 to 45 months	Yes	40	7.7%
between 41 to 45 months	No	478	92.3%
between 46 to 50 months	Yes	50	10.0%
between 46 to 50 months	No	449	89.8%
between 51 to 55 months	Yes	72	13.4%
between 51 to 55 months	No	463	86.4%
above 56	Yes	25	4.9%
above 56	No	482	94.9%

```
chisq.test(df$age_in_months_cat , df$EC6)
```

```
## Warning in chisq.test(df$age_in_months_cat, df$EC6): Chi-squared
approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC6
## X-squared = 47.66, df = 12, p-value = 3.581e-06
```

```
ggplot(tab1, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC6)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Can the child identify or name at least ten letters of the
alphabet?" , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```




```
#-----------------------------------#
```

```
tab2 = df %>% group_by( age_in_months_cat , EC7) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))
```

```
## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.
```

```
tab2 = tab2 [(tab2$EC7 == 1  | tab2$EC7 ==2),]
tab2$EC6 [tab2$EC7 ==1] = "Yes"
```

```
## Warning: Unknown or uninitialised column: `EC6`.
```

```
tab2$EC6 [tab2$EC7 ==2] = "No"
tab2 = tab2[!is.na(tab2$age_in_months_cat),]
kable(tab2 , format = "markdown")
```

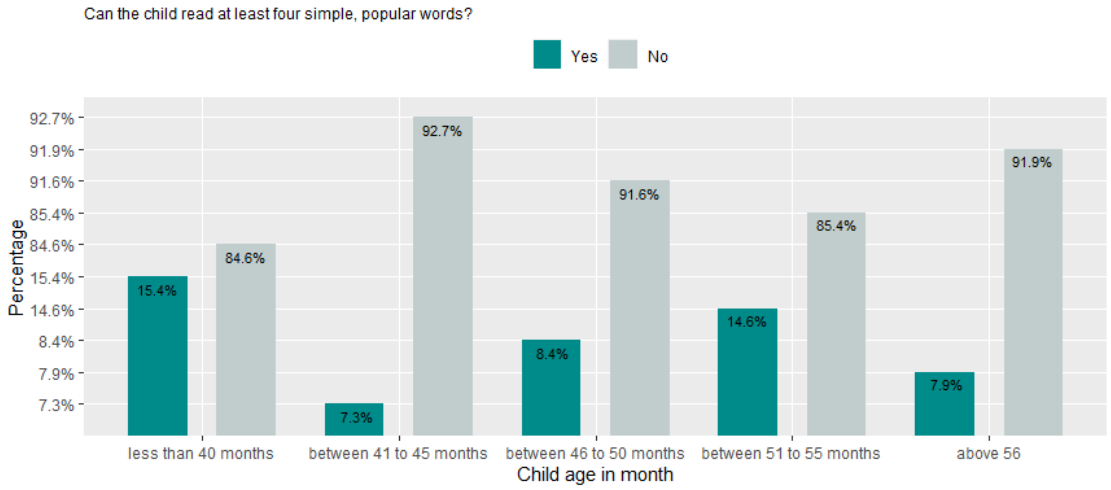
age_in_months_cat	EC7	n	perc	EC6
less than 40 months	1	66	15.4%	Yes
less than 40 months	2	363	84.6%	No
between 41 to 45 months	1	38	7.3%	Yes
between 41 to 45 months	2	480	92.7%	No
between 46 to 50 months	1	42	8.4%	Yes
between 46 to 50 months	2	458	91.6%	No
between 51 to 55 months	1	78	14.6%	Yes
between 51 to 55 months	2	458	85.4%	No
above 56	1	40	7.9%	Yes
above 56	2	467	91.9%	No

```
chisq.test(df$age_in_months_cat , df$EC7)
```

```
## Warning in chisq.test(df$age_in_months_cat, df$EC7): Chi-squared
approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC7
## X-squared = 35.41, df = 8, p-value = 2.251e-05
```

```
ggplot(tab2, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC7)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Can the child read at least four simple, popular words?"  , fill =
"") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```




```
#-------------------------------#

tab3 = df %>% group_by( age_in_months_cat , EC8) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))

## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.

tab3 = tab3 [(tab3$EC8 == 1  | tab3$EC8 ==2),]
tab3$EC8 [tab3$EC8 ==1] = "Yes"
tab3$EC8 [tab3$EC8 ==2] = "No"
tab3 = tab3[!is.na(tab3$age_in_months_cat),]
kable(tab3 , format = "markdown")
```

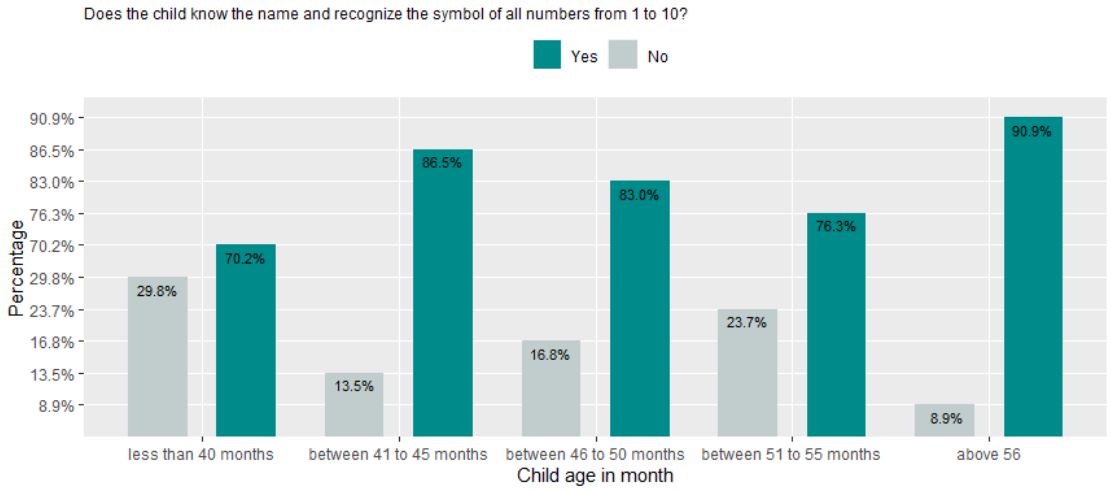
age_in_months_cat	EC8	n	perc
less than 40 months	Yes	128	29.8%
less than 40 months	No	301	70.2%
between 41 to 45 months	Yes	70	13.5%
between 41 to 45 months	No	448	86.5%
between 46 to 50 months	Yes	84	16.8%
between 46 to 50 months	No	415	83.0%
between 51 to 55 months	Yes	127	23.7%
between 51 to 55 months	No	409	76.3%
above 56	Yes	45	8.9%
above 56	No	462	90.9%

```
chisq.test(df$age_in_months_cat , df$EC8)

## Warning in chisq.test(df$age_in_months_cat, df$EC8): Chi-squared
approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC8
## X-squared = 95.509, df = 12, p-value = 4.2e-15
```

```
ggplot(tab3, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC8)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Does the child know the name and recognize the symbol of all numbers
from 1 to 10?"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```



Physical capacity

```
# EC9: "Can (name) pick up a small object with two fingers, like a stick or a
rock from the ground?" "Yes=1/No=2/DK=8"
# EC10: "Is (name) sometimes too sick to play?" "Yes=1/No=2/DK=8"
```

```
tab4 = df %>% group_by( age_in_months_cat , EC9) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))
```

```
## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.
```

```
tab4 = tab4 [(tab4$EC9 == 1  | tab4$EC9 ==2),]
tab4$EC9 [tab4$EC9 ==1] = "Yes"
tab4$EC9 [tab4$EC9 ==2] = "No"
tab4 = tab4[!is.na(tab4$age_in_months_cat),]
kable(tab4 , format = "markdown")
```

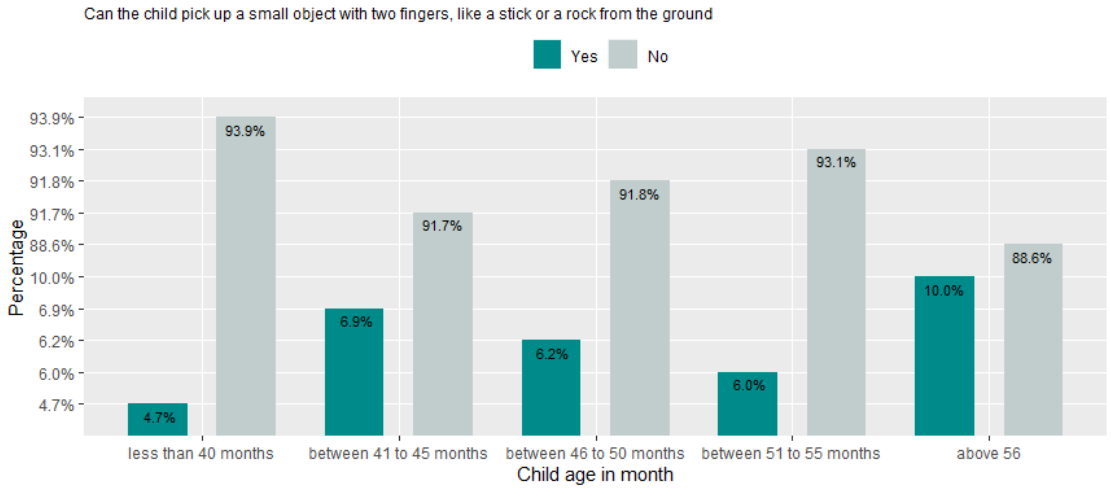
age_in_months_cat	EC9	n	perc
less than 40 months	Yes	403	93.9%
less than 40 months	No	20	4.7%
between 41 to 45 months	Yes	475	91.7%
between 41 to 45 months	No	36	6.9%
between 46 to 50 months	Yes	459	91.8%
between 46 to 50 months	No	31	6.2%
between 51 to 55 months	Yes	499	93.1%
between 51 to 55 months	No	32	6.0%
above 56	Yes	450	88.6%
above 56	No	51	10.0%

```
chisq.test(df$age_in_months_cat , df$EC9)
```

```
## Warning in chisq.test(df$age_in_months_cat, df$EC9): Chi-squared
approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC9
## X-squared = 18.634, df = 12, p-value = 0.09776
```

```
ggplot(tab4, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC9)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Can the child pick up a small object with two fingers, like a stick
or a rock from the ground"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```




```
#--------------------------------------#


tab5 = df %>% group_by( age_in_months_cat , EC10) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))

## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.

tab5 = tab5 [(tab5$EC10 == 1  | tab5$EC10 ==2),]
tab5$EC10 [tab5$EC10 ==1] = "Yes"
tab5$EC10 [tab5$EC10 ==2] = "No"
tab5 = tab5[!is.na(tab5$age_in_months_cat),]
kable(tab5 , format = "markdown")
```

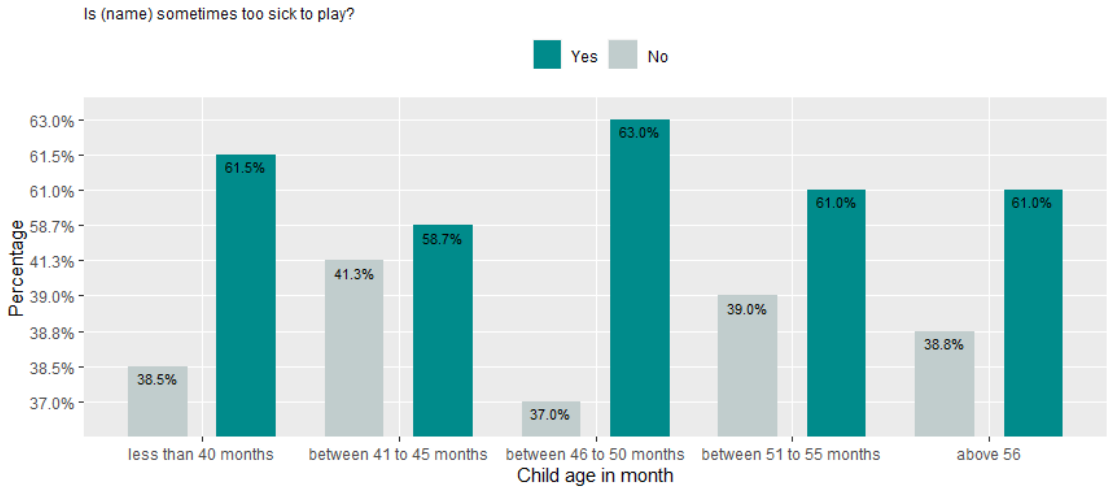
age_in_months_cat	EC10	n	perc
less than 40 months	Yes	165	38.5%
less than 40 months	No	264	61.5%
between 41 to 45 months	Yes	214	41.3%
between 41 to 45 months	No	304	58.7%
between 46 to 50 months	Yes	185	37.0%
between 46 to 50 months	No	315	63.0%
between 51 to 55 months	Yes	209	39.0%
between 51 to 55 months	No	327	61.0%
above 56	Yes	197	38.8%
above 56	No	310	61.0%

```
chisq.test(df$age_in_months_cat , df$EC10)
```

```
## Warning in chisq.test(df$age_in_months_cat, df$EC10): Chi-squared
approximation
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC10
## X-squared = 5.9668, df = 8, p-value = 0.651
```

```
ggplot(tab5, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC10)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Is (name) sometimes too sick to play?"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```



Learning capacity

```
tab6 = df %>% group_by( age_in_months_cat , EC11) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))
```

```
## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.
```

```
tab6 = tab6 [(tab6$EC11 == 1  | tab6$EC11 ==2),]
tab6$EC11 [tab6$EC11 ==1] = "Yes"
tab6$EC11 [tab6$EC11 ==2] = "No"
tab6 = tab6[!is.na(tab6$age_in_months_cat),]
kable(tab6 , format = "markdown")
```

age_in_months_cat	EC11	n	perc

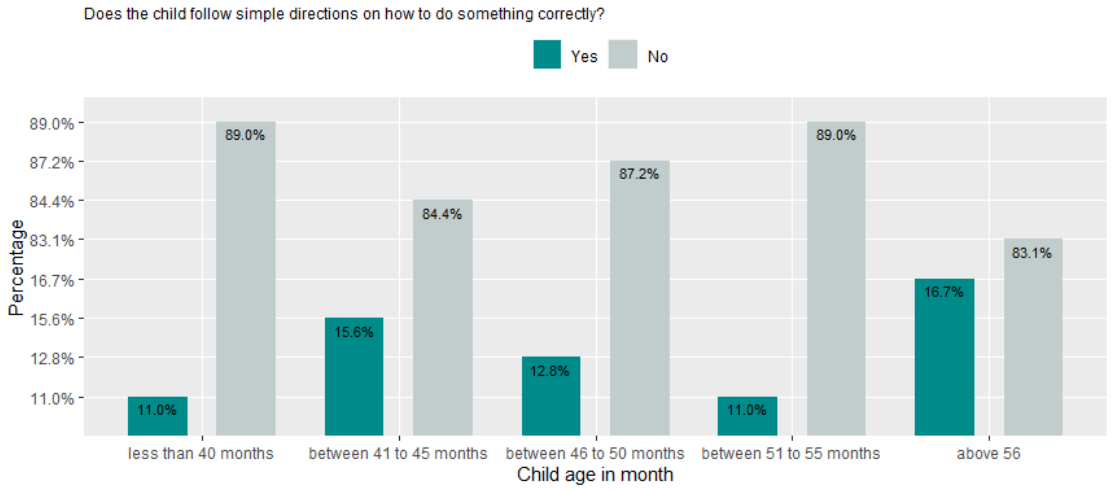
age_in_months_cat	EC11	n	perc
less than 40 months	Yes	382	89.0%
less than 40 months	No	47	11.0%
between 41 to 45 months	Yes	437	84.4%
between 41 to 45 months	No	81	15.6%
between 46 to 50 months	Yes	436	87.2%
between 46 to 50 months	No	64	12.8%
between 51 to 55 months	Yes	477	89.0%
between 51 to 55 months	No	59	11.0%
above 56	Yes	422	83.1%
above 56	No	85	16.7%

```
chisq.test(df$age_in_months_cat , df$EC11)
```

```
## Warning in chisq.test(df$age_in_months_cat, df$EC11): Chi-squared
approximation
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC11
## X-squared = 16, df = 8, p-value = 0.04238
```

```
ggplot(tab6, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC11)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Does the child follow simple directions on how to do something
correctly?"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```




```
#-----------------------------------------------#

tab7 = df %>% group_by( age_in_months_cat , EC12) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))

## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.

tab7 = tab7 [(tab7$EC12 == 1  | tab7$EC12 ==2),]
tab7$EC12 [tab7$EC12 ==1] = "Yes"
tab7$EC12 [tab7$EC12 ==2] = "No"
tab7 = tab7[!is.na(tab7$age_in_months_cat),]
kable(tab7 , format = "markdown")
```

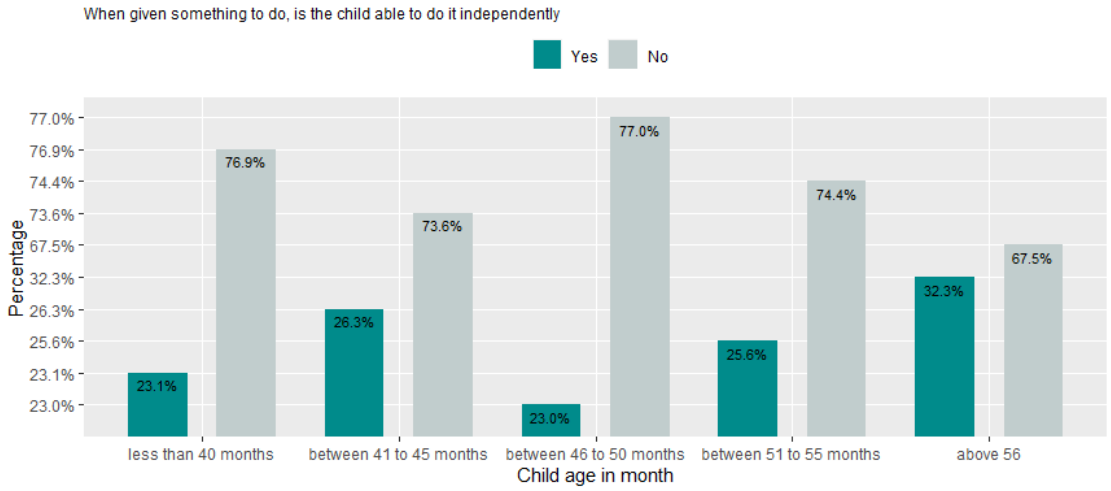
age_in_months_cat	EC12	n	perc
less than 40 months	Yes	330	76.9%
less than 40 months	No	99	23.1%
between 41 to 45 months	Yes	381	73.6%
between 41 to 45 months	No	136	26.3%
between 46 to 50 months	Yes	385	77.0%
between 46 to 50 months	No	115	23.0%
between 51 to 55 months	Yes	399	74.4%
between 51 to 55 months	No	137	25.6%
above 56	Yes	343	67.5%
above 56	No	164	32.3%

```
chisq.test(df$age_in_months_cat , df$EC12)

## Warning in chisq.test(df$age_in_months_cat, df$EC12): Chi-squared
approximation
## may be incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC12
## X-squared = 22.558, df = 12, p-value = 0.03173
```

```
ggplot(tab7, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC12)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "When given something to do, is the child able to do it
independently"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```



Socio-emotional skills

```
tab8 = df %>% group_by( age_in_months_cat , EC13) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))
```

```
## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.
```

```
tab8 = tab8 [(tab8$EC13 == 1  | tab8$EC13 ==2),]
tab8$EC13 [tab8$EC13 ==1] = "Yes"
tab8$EC13 [tab8$EC13 ==2] = "No"
tab8 = tab8[!is.na(tab8$age_in_months_cat),]
kable(tab8 , format = "markdown")
```

age_in_months_cat	EC13	n	perc
less than 40 months	Yes	412	96.0%
less than 40 months	No	17	4.0%

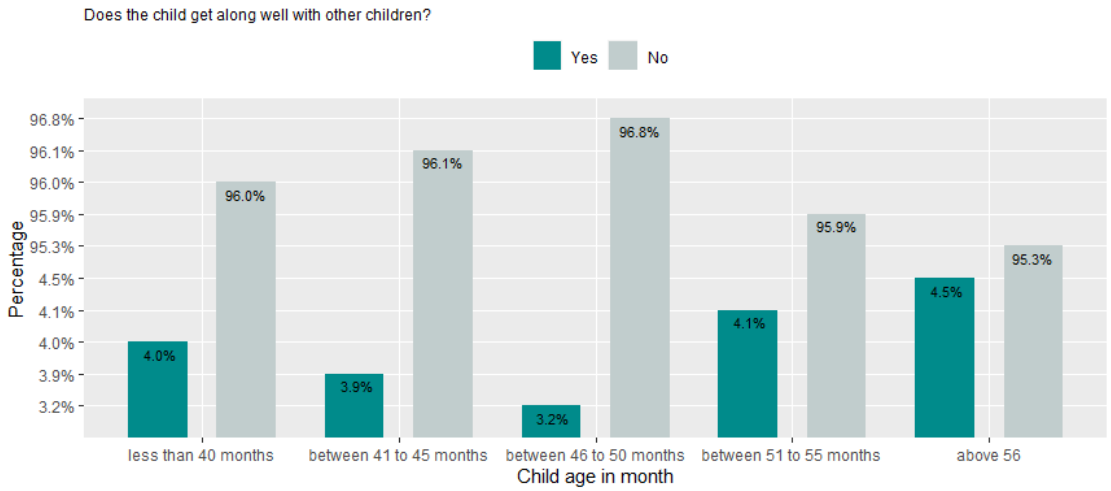
age_in_months_cat	EC13	n	perc
between 41 to 45 months	Yes	498	96.1%
between 41 to 45 months	No	20	3.9%
between 46 to 50 months	Yes	484	96.8%
between 46 to 50 months	No	16	3.2%
between 51 to 55 months	Yes	514	95.9%
between 51 to 55 months	No	22	4.1%
above 56	Yes	484	95.3%
above 56	No	23	4.5%

```
chisq.test(df$age_in_months_cat , df$EC13)

## Warning in chisq.test(df$age_in_months_cat, df$EC13): Chi-squared
approximation
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC13
## X-squared = 5.1533, df = 8, p-value = 0.7411

ggplot(tab8, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC13)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Does the child get along well with other children?"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```



```
#-------------------------------------#

tab9 = df %>% group_by( age_in_months_cat , EC14) %>% summarise(n = n() ) %>%
dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))

## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.

tab9 = tab9 [(tab9$EC14 == 1  | tab9$EC14 ==2),]
tab9$EC14 [tab9$EC14 ==1] = "Yes"
tab9$EC14 [tab9$EC14 ==2] = "No"
tab9 = tab9[!is.na(tab9$age_in_months_cat),]
kable(tab9 , format = "markdown")
```

age_in_months_cat	EC14	n	perc
less than 40 months	Yes	190	44.3%
less than 40 months	No	239	55.7%
between 41 to 45 months	Yes	256	49.4%
between 41 to 45 months	No	261	50.4%
between 46 to 50 months	Yes	249	49.8%
between 46 to 50 months	No	251	50.2%
between 51 to 55 months	Yes	254	47.4%
between 51 to 55 months	No	281	52.4%
above 56	Yes	270	53.1%
above 56	No	236	46.5%

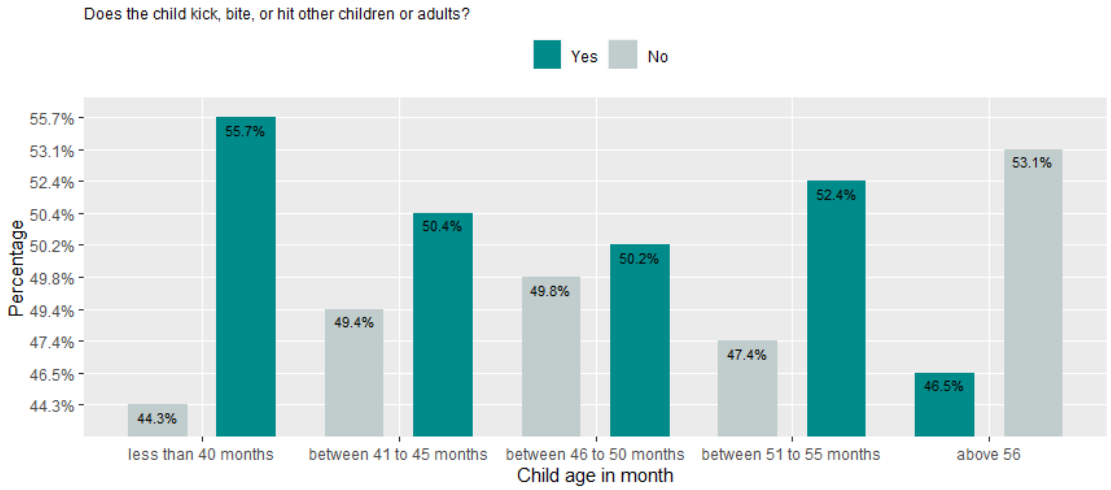
```
chisq.test(df$age_in_months_cat , df$EC14)

## Warning in chisq.test(df$age_in_months_cat, df$EC14): Chi-squared
approximation
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC14
## X-squared = 14.031, df = 12, p-value = 0.2987

ggplot(tab9, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC14)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Does the child kick, bite, or hit other children or adults?"  , fill
= "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
```

```
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```




```
#-----------------------------------------#


tab10 = df %>% group_by( age_in_months_cat , EC15) %>% summarise(n = n() )
%>% dplyr::mutate(perc = paste0(sprintf("%4.1f", n / sum(n) * 100), "%"))

## `summarise()` has grouped output by 'age_in_months_cat'. You can override
using
## the `.groups` argument.

tab10 = tab10 [(tab10$EC15 == 1  | tab10$EC15 ==2),]
tab10$EC15 [tab10$EC15 ==1] = "Yes"
tab10$EC15 [tab10$EC15 ==2] = "No"
tab10 = tab10[!is.na(tab10$age_in_months_cat),]
kable(tab10 , format = "markdown")
```

age_in_months_cat	EC15	n	perc
less than 40 months	Yes	158	36.8%
less than 40 months	No	271	63.2%
between 41 to 45 months	Yes	200	38.6%
between 41 to 45 months	No	315	60.8%
between 46 to 50 months	Yes	195	39.0%
between 46 to 50 months	No	304	60.8%
between 51 to 55 months	Yes	188	35.1%
between 51 to 55 months	No	348	64.9%
above 56	Yes	190	37.4%
above 56	No	317	62.4%

```
chisq.test(df$age_in_months_cat , df$EC15)
```

```
## Warning in chisq.test(df$age_in_months_cat, df$EC15): Chi-squared
approximation
## may be incorrect

##
##  Pearson's Chi-squared test
##
## data:  df$age_in_months_cat and df$EC15
## X-squared = 14.33, df = 12, p-value = 0.2802
```

```
ggplot(tab10, aes(x = age_in_months_cat, y = perc, fill = as.factor (EC15)))+
geom_bar(stat = "identity", position=position_dodge(.9) , width = 0.6) +
geom_text(aes(label = perc), vjust = 1.5, colour = "black" , size = 3,
position=position_dodge(.9))+ #text labels
labs(x = "Child age in month", y = "Percentage",
title = "Does the child get distracted easily"  , fill = "") +
scale_fill_manual(values = c("darkcyan", "azure3"),
                    labels = c("Yes", "No")) +   theme(legend.position =
"top" , plot.title = element_text(size = 9))
```