Women Who Exercise

Torrey Capobianco
Bellevue University

Exploratory Data Analysis
Fall 2019

Hypothesis

Women who have children under the age of 18 exercise less than women who do not have children under the age of 18.

The Data

American Time Use Survey
2003-2015
US Bureau of Labor Statistics

Exercise Data

Who?

This analysis looks at exercise durations of females recorded in a single day.

What is considered exercise?

Durations recorded were taken from category “Participating in Sports, Exercise, and Recreation.” Due to this analysis looking at higher pace exercise activities, recreation activities such as playing billiards, fishing, hunting, vehicle touring/racing were removed from the data.

Variable Selection

Classification Variables

tucaseid:

case ID for respondent

tesex:

sex of respondent

trtier1p:

activity category 1st hierarchy

trtier2p:

activity category 2nd hierarchy

trcodep:

6 digit activity code

Analysis Variables

trchildnum:

number of children for respondent

tehruslt:

usual hours worked per week for respondent

tespuhrs:

usual hours worked per week for respondent's spouse

teage:

age of respondent

tuactdur24:

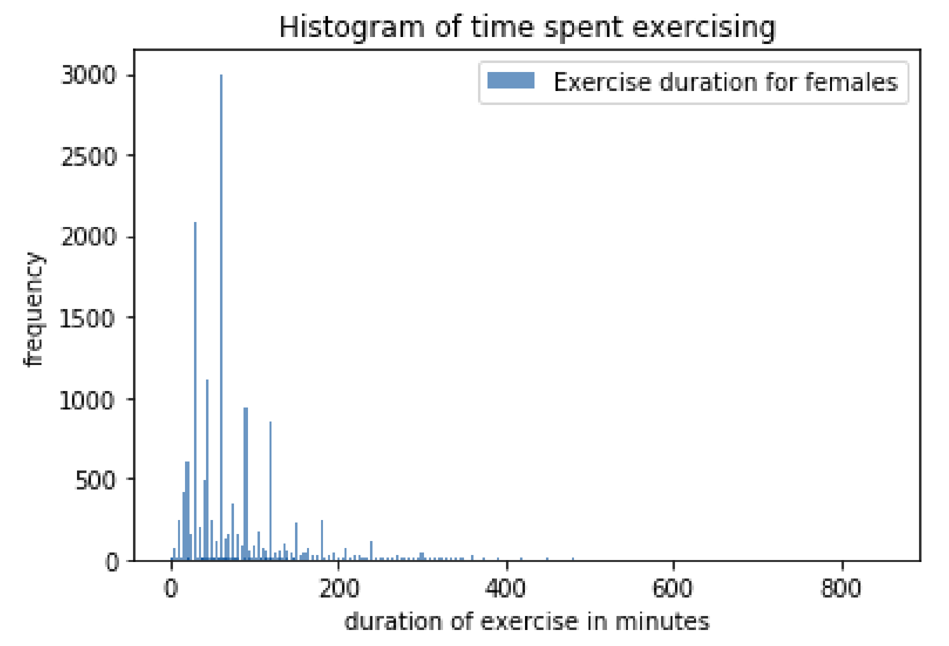
duration of activity in minutes

Histograms

Histograms and descriptive characteristics

Time spent exercising

tuactdur24



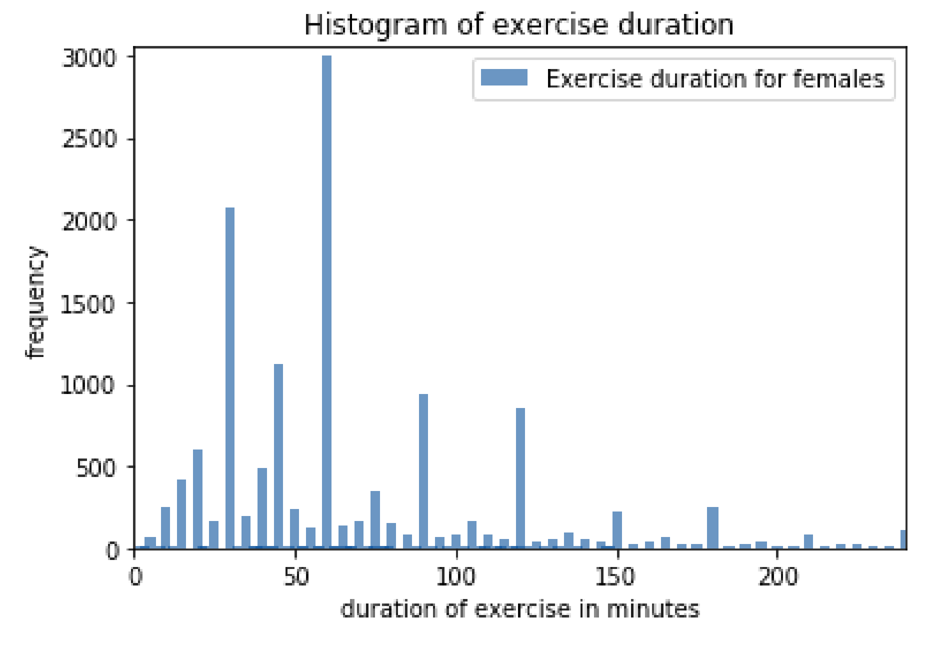
Mean:	77.38 minutes
Mode:	60 minutes
Variance:	4676.86
Standard Deviation:	68.37

The histogram is left skewed with most of the time spent exercising in the left tail with the right tail extending far to the right.

Outliers: There are cases of recorded exercise up to 15 hours in a day. This seems unreasonable. For this analysis all records greater than 240 minutes or 4 hours is removed.

Time spent exercising

tuactdur24 – removed outliers

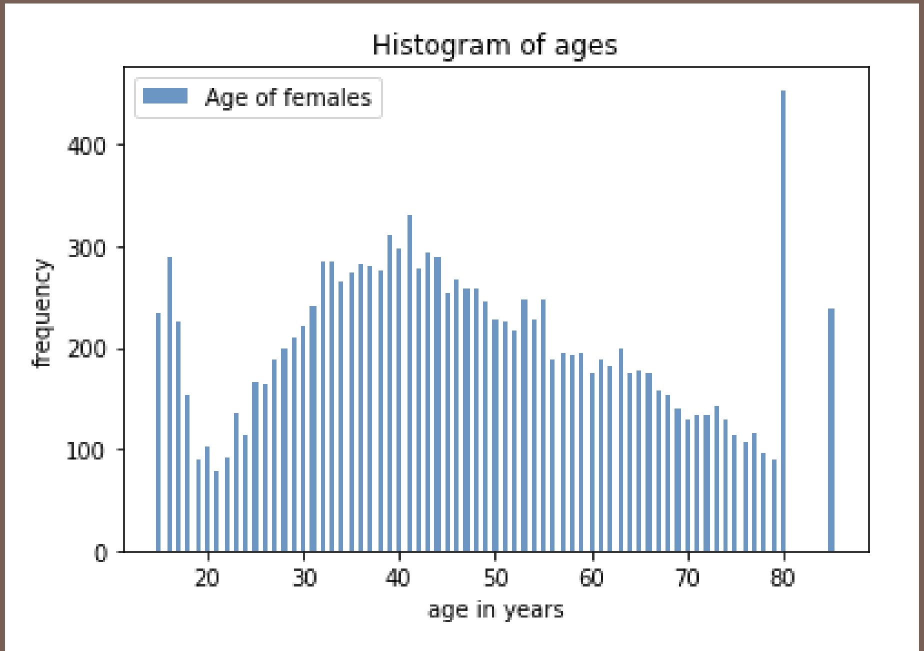


Mean:	68.51 minutes
Mode:	60 minutes
Variance:	2198.39
Standard Deviation:	46.89

The histogram is left skewed with most of the time spent exercising in the left tail with the right tail extending far to the right.

Age

teage

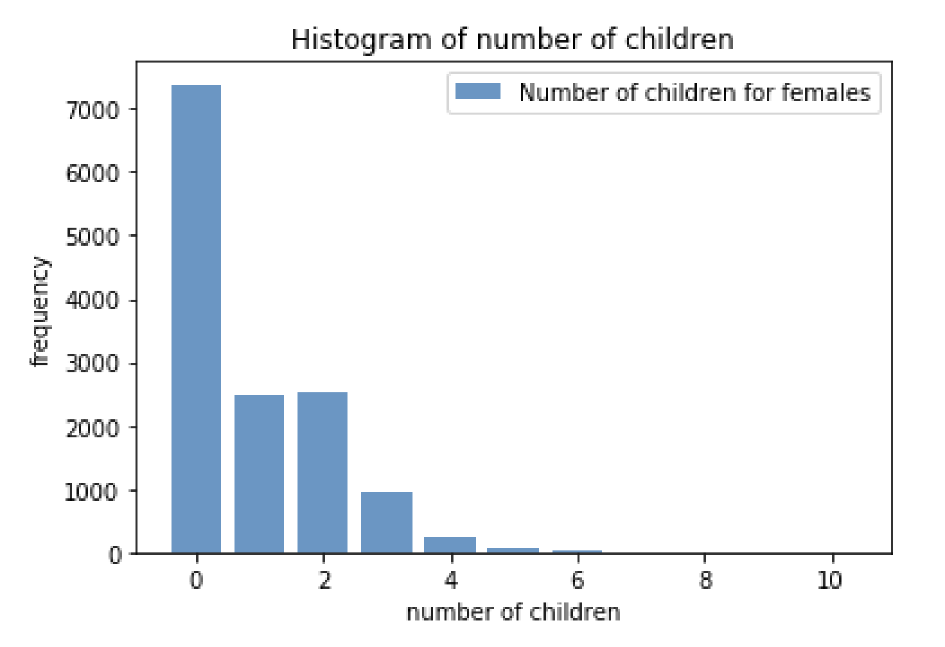


Mean:	46.96 years
Mode:	80 years
Variance:	326.58
Standard Deviation:	18.07

The histogram is multimodal, with a range with females less than their 20's, mid 40's, and the right tail their 80's.

Number of children

trchildnum

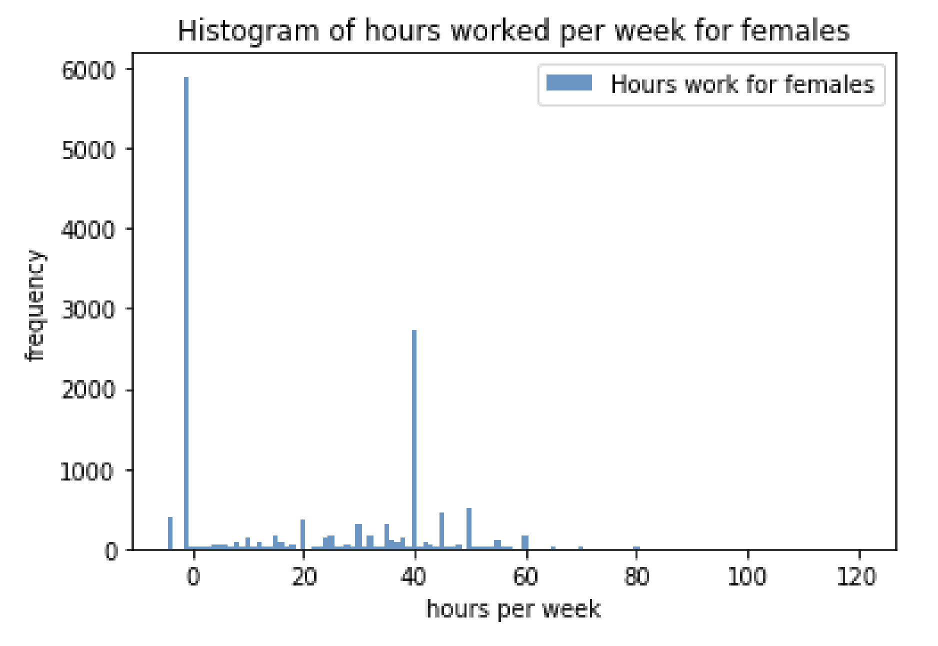


Mean:	0.87 children
Mode:	0 children
Variance:	1.28
Standard Deviation:	1.13

The histogram is left negative skewed with a tail extending far to the right.

Usual hours work per week

tehruslt



Mean:	19.05 hours
Mode:	-1 (does not work)
Variance:	443.08
Standard Deviation:	21.05

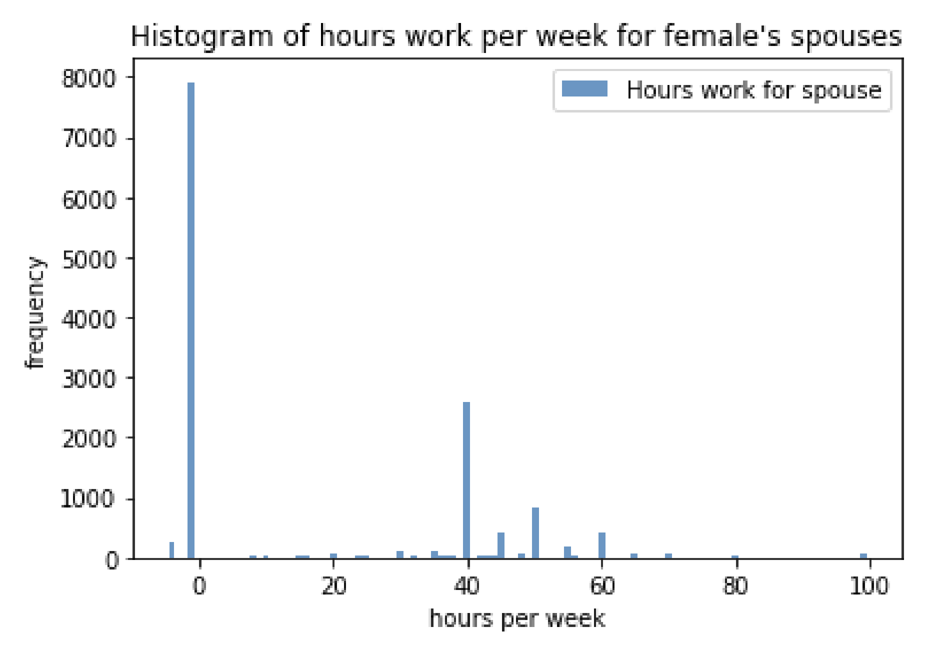
The histogram is binominal, with two peaks at -1 (does not work) and 40 hours.

Respondent answer of -4 indicates hours vary.

Respondent answer of -1 indicates does not work.

Usual hours work per week for spouse

tespuhrs



Mean:	17.33 hours
Mode:	-1 (does not work)
Variance:	563.63
Standard Deviation:	23.74

The histogram is binominal, with two peaks at -1 (does not work) and 40 hours.

Respondent answer of -4 indicates hours vary.

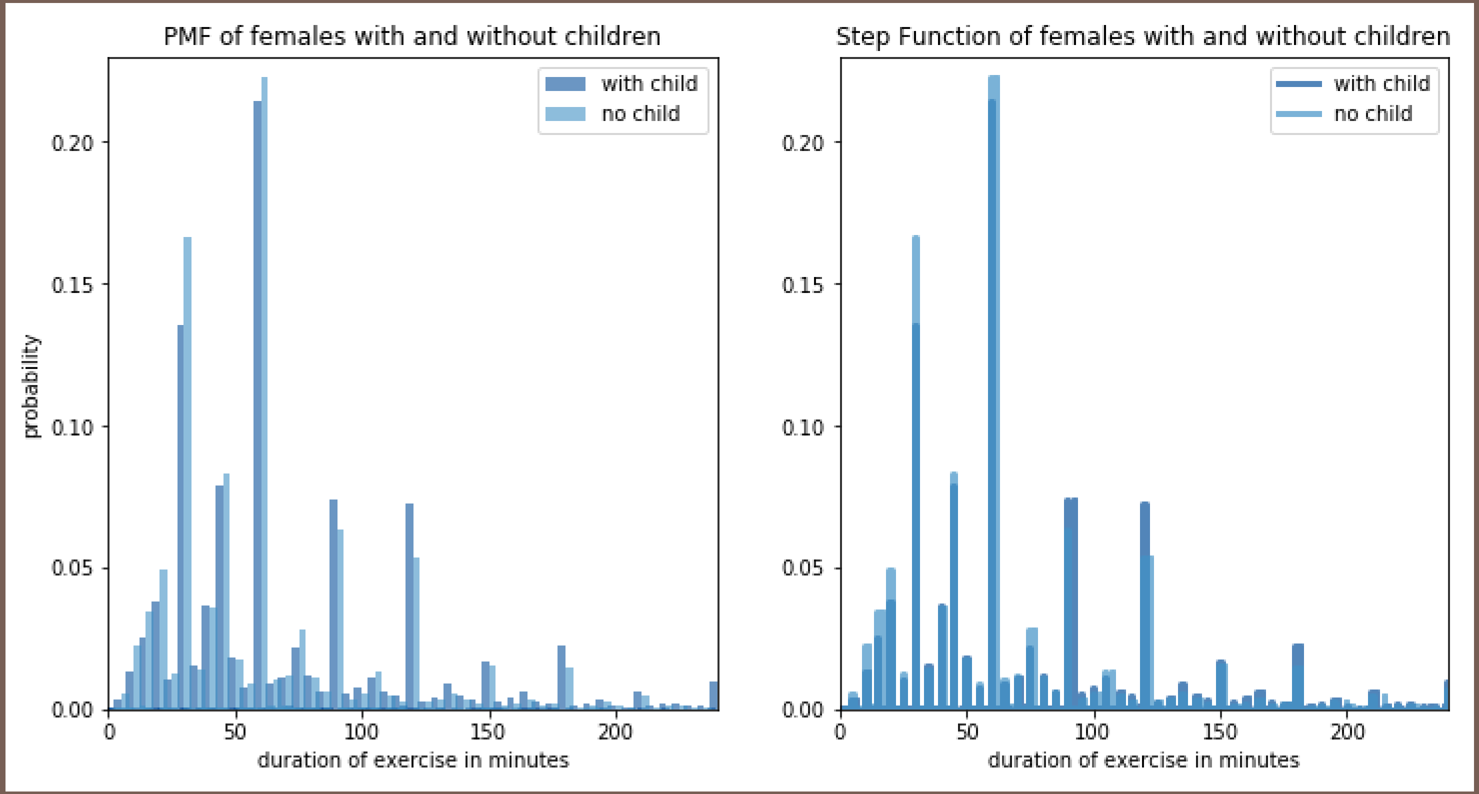
Respondent answer of -1 indicates does not work.

Probability Mass Function

Comparing two scenarios with PMF

PMF of female exercise duration

Females with and without children



PMF of females with and without children

Step Function of females with and without children

Exercise Duration	With Child	No Child
Mean	73.34	64.36
Variance	2401.37	1987.65
Standard Deviation	49.00	44.58

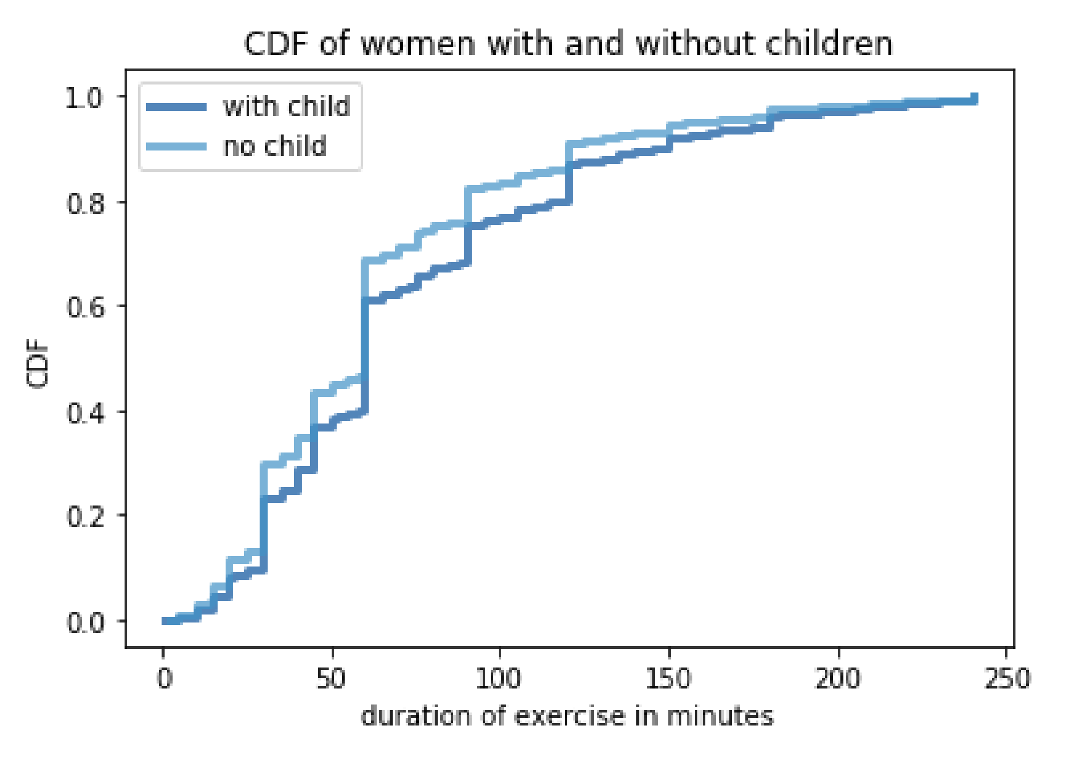
The probability that females with children exercising for shorter durations is smaller than those without children. Compared to longer durations, they have a higher probability than those without children.

Cumulative Distribution Function

Comparing two scenarios with CDF

CDF of female exercise duration

Females with and without children



Exercise Duration	With Child	No Child
≤ 100 minutes	78 %	82 %
≤ 60 minutes	60 %	60 %
≤ 50 minutes	39 %	45 %
≤ 30 minutes	25 %	25 %

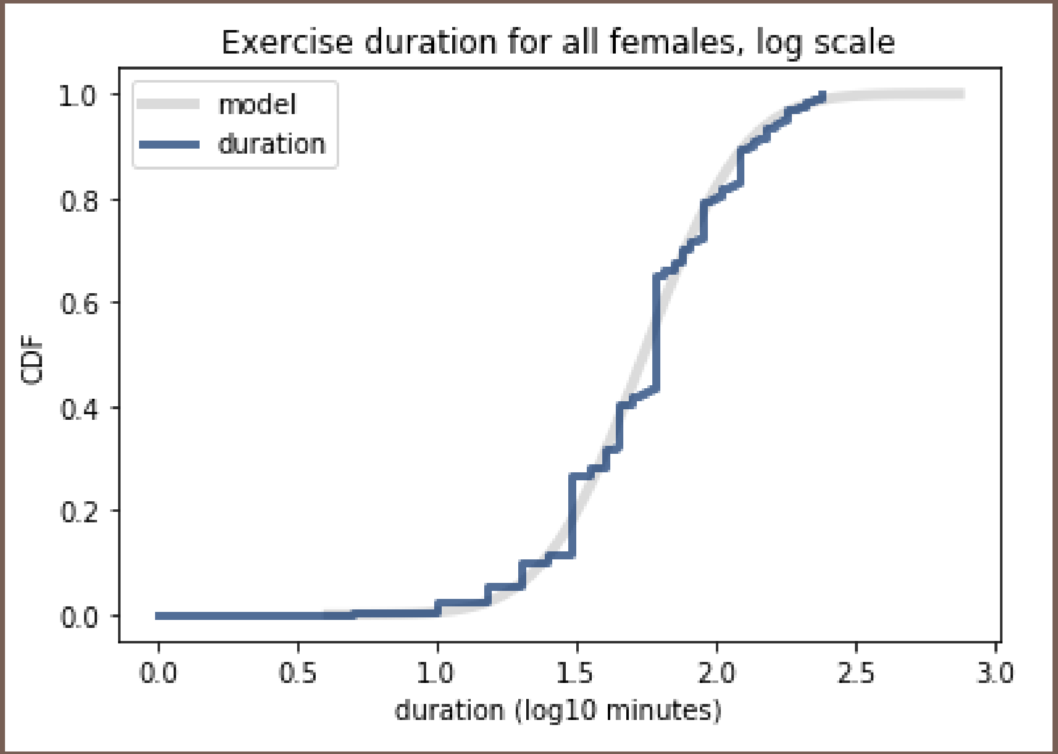
The CDF of exercise duration for females with children and no children are common in some durations. Once past 60 minutes, females without children have a higher cumulative probability of exercising more.

Analytical Distribution

Lognormal distribution

Lognormal Distribution

Exercise duration for all females



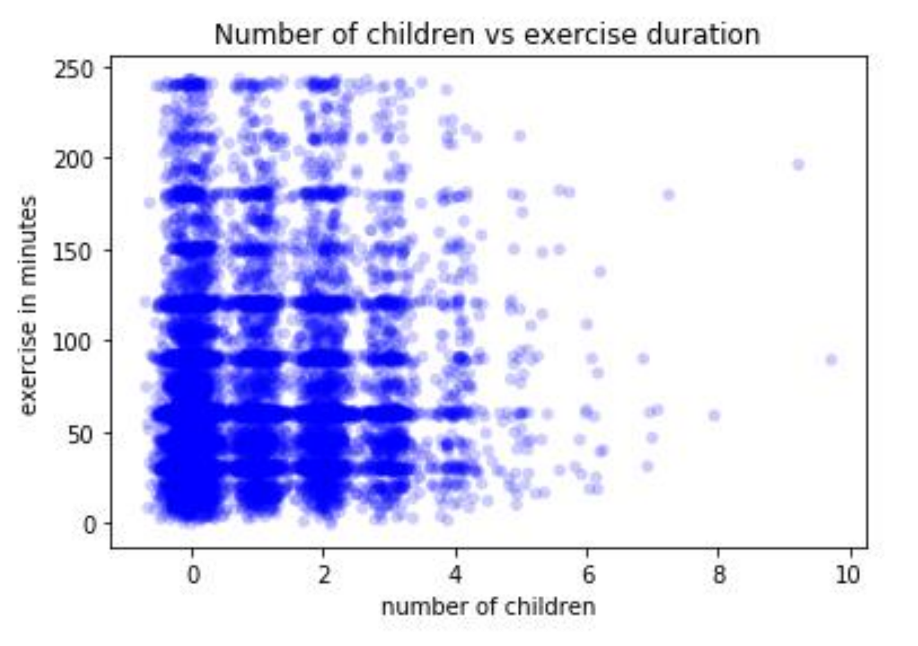
Lognormal distribution is the best fit for modeling the exercise duration of females.

Variable Relationships

Scatter plots, covariance, and correlation.

Number of children vs Exercise duration

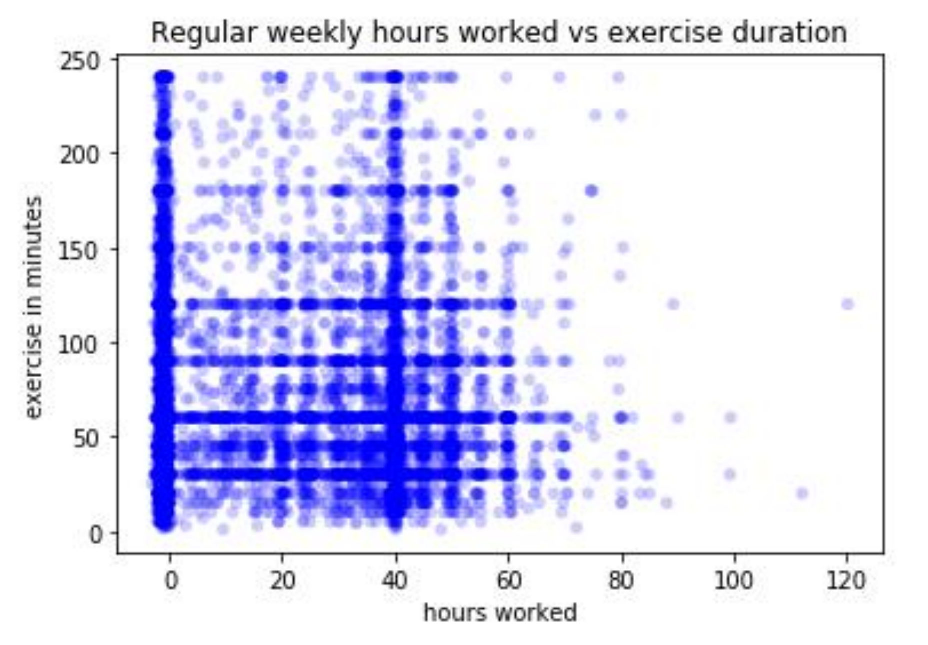
Scatter Plot



	Result	Interpretation
Covariance	4.54	Positive, variables move in the same direction
Pearson’s Correlation	0.08	Close to zero, no relationship between variables
Spearman’s Rank Correlation	0.08	Close to zero, no relationship between variables

Number of hours work vs Exercise duration

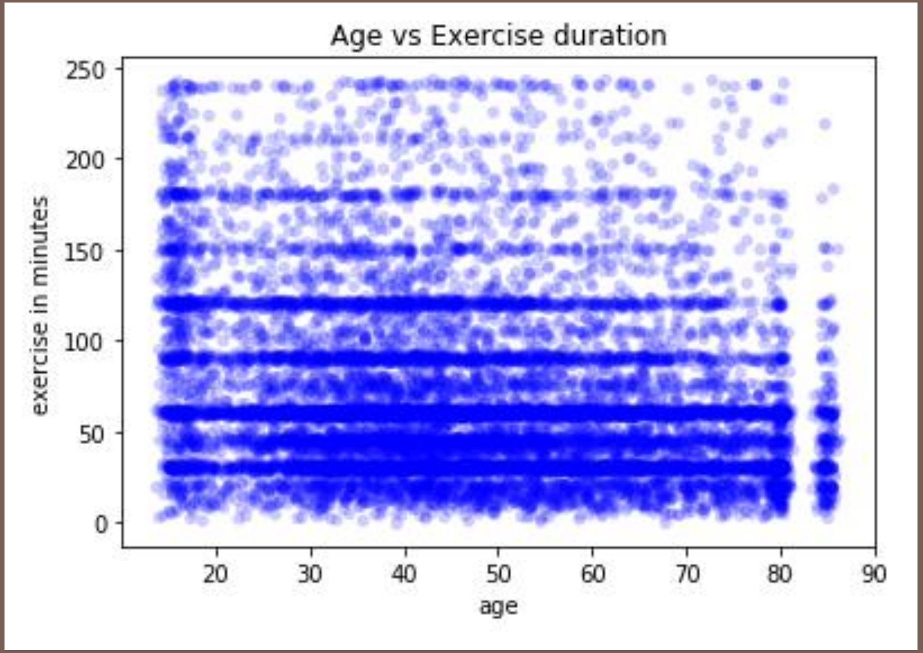
Scatter Plot



	Result	Interpretation
Covariance	-8.33	Negative, variables move inversely
Pearson’s Correlation	-0.008	Close to zero, no relationship between variables
Spearman’s Rank Correlation	0.003	Close to zero, no relationship between variables

Age vs Exercise duration

Scatter Plot



	Result	Interpretation
Covariance	-160.5	Negative, variables move inversely
Pearson’s Correlation	-0.19	Close to zero, no relationship between variables
Spearman’s Rank Correlation	-0.19	Close to zero, no relationship between variables

Hypothesis Test

Permutation difference of the means.

Women with children exercise less than those without

Null Hypothesis:

There is no difference in exercise habits between women with children under the age of 18 and women without children.

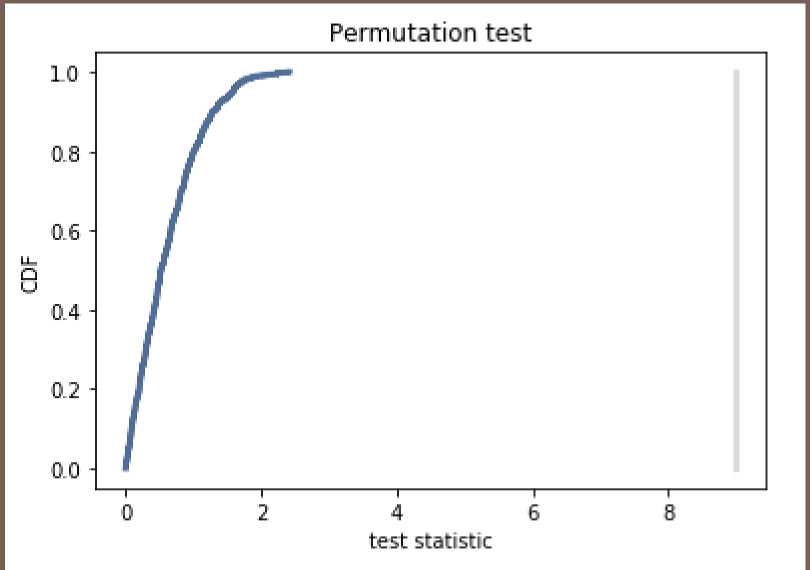
Permutation Difference of the Means:

P-Value: 0.0

Conclusion:

P-value is less than 1%. The effect is likely not due to chance.

- ➢ Reject the null hypothesis
- ➢ Accept the alternative hypothesis

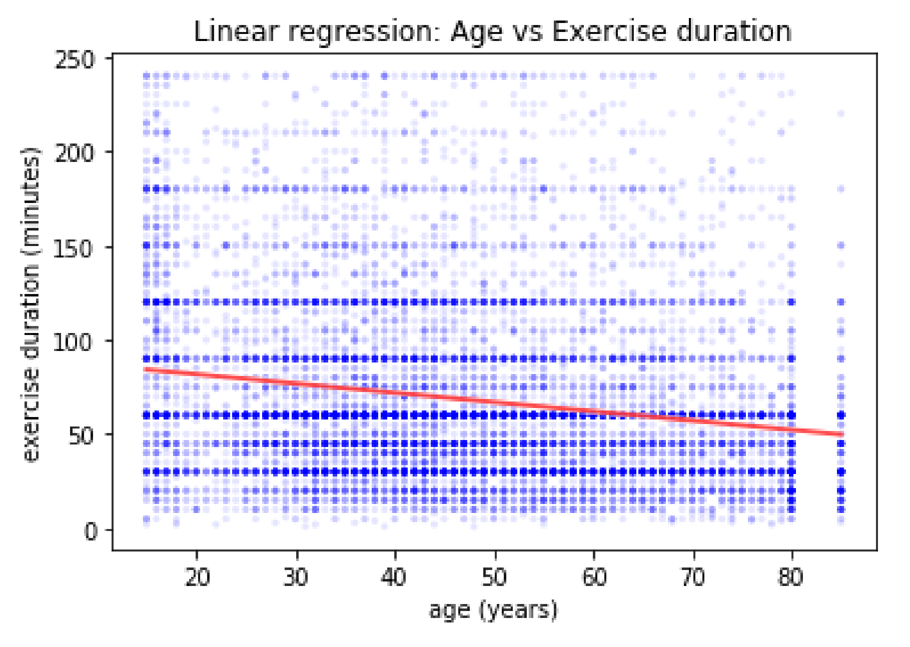


Regression

Simple and multiple regression.

Simple Regression

Linear Least Squares:



Exercise duration as a function of age:

r-squared: 0.036

p-value: 0.00

Age accounts for 3.6% of exercise duration in females.

Simple Regression

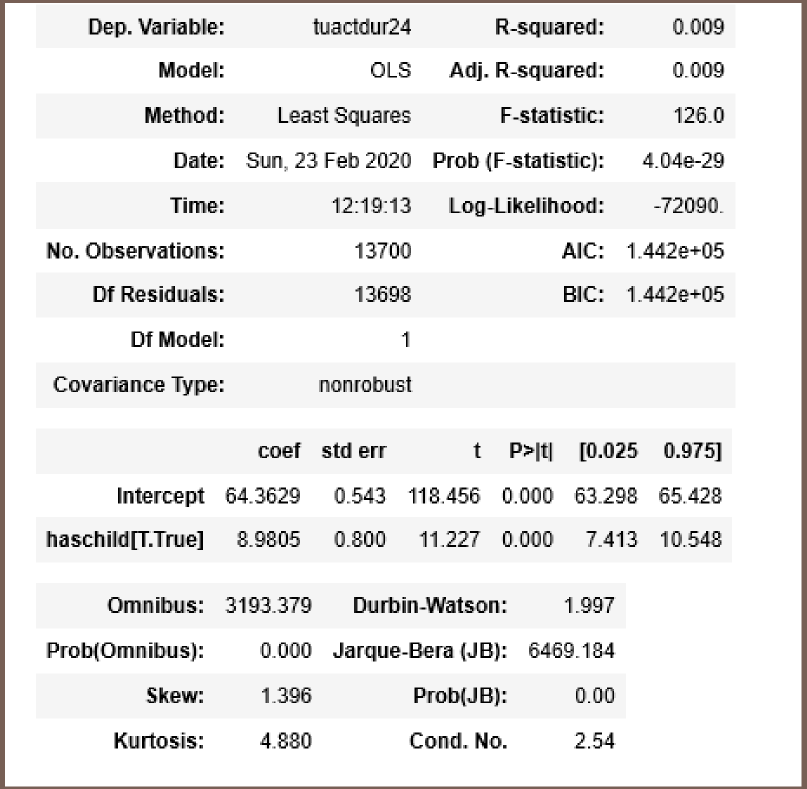
Exercise duration as a function of having a child:

r-squared: 0.009

Having a child accounts for 0.9% of exercise duration in females.

p-value: 0.00

This effect is likely to not occur by chance.



Dep. Variable:	tuactdur24	R-squared:	0.009
Model:	OLS	Adj. R-squared:	0.009
Method:	Least Squares	F-statistic:	126.0
Date:	Sun, 23 Feb 2020	Prob (F-statistic):	4.04e-29
Time:	12:19:13	Log-Likelihood:	-72090.
No. Observations:	13700	AIC:	1.442e+05
Df Residuals:	13698	BIC:	1.442e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	64.3629	0.543	118.456	0.000	63.298	65.428
haschild[T.True]	8.9805	0.800	11.227	0.000	7.413	10.548

Omnibus:	3193.379	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	6469.184
Skew:	1.396	Prob(JB):	0.00
Kurtosis:	4.880	Cond. No.	2.54

Multiple Regression

Exercise duration as a function of having a child and age:

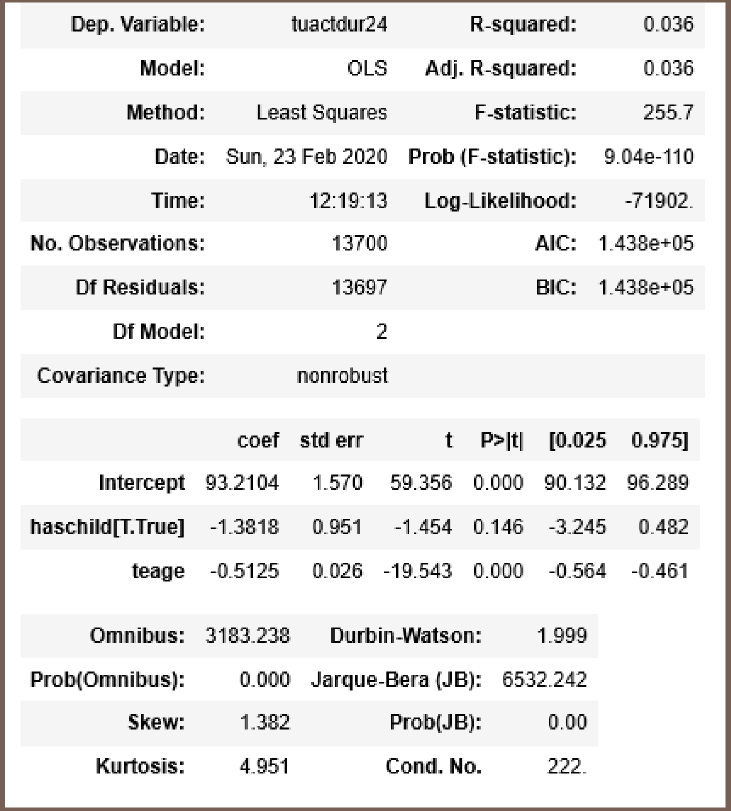
r-squared: 0.036

r-squared increased to 3.6% accountability with adding age.

p-value has child: 0.146 p-value age: 0.00

The effect of age is likely to not occur by chance.

Having a child > .05 results not being statistically significant. This does not have an affect in determining exercise when age is involved.



```
    Dep. Variable:        tuactdur24        R-squared:      0.036
            Model:               OLS    Adj. R-squared:      0.036
           Method:     Least Squares       F-statistic:      255.7
             Date:  Sun, 23 Feb 2020 Prob (F-statistic):  9.04e-110
             Time:          12:19:13    Log-Likelihood:    -71902.
No. Observations:             13700               AIC:  1.438e+05
     Df Residuals:             13697               BIC:  1.438e+05
         Df Model:                 2
  Covariance Type:         nonrobust

                    coef  std err        t  P>|t|  [0.025  0.975]
       Intercept  93.2104    1.570   59.356  0.000  90.132  96.289
 haschild[T.True] -1.3818    0.951   -1.454  0.146  -3.245   0.482
            teage -0.5125    0.026  -19.543  0.000  -0.564  -0.461

       Omnibus: 3183.238     Durbin-Watson:     1.999
 Prob(Omnibus):    0.000  Jarque-Bera (JB):  6532.242
          Skew:    1.382          Prob(JB):      0.00
      Kurtosis:    4.951          Cond. No.      222.
```

Multiple Regression

Exercise duration as a function of having children, age, hours work per week for self and for spouse:

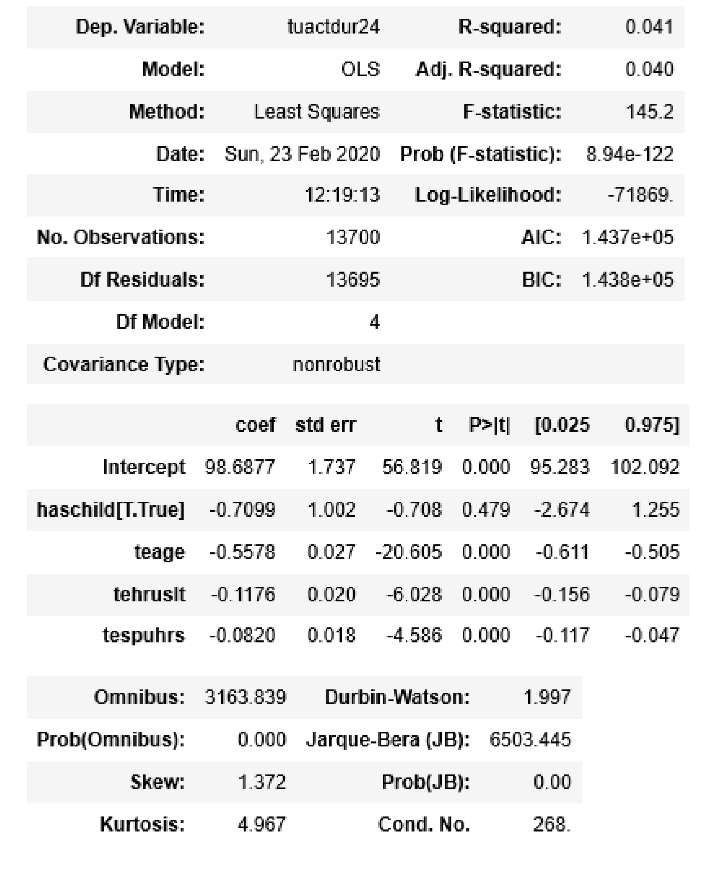
r-squared: 0.041

r-squared increased to 4.1% accountability with adding hours work per week for self and spouse.

p-value has child: 0.479

p-value all others: 0.00

The effect of all others but having a child are likely to not occur by chance. Having a child has no affect on exercise duration with a p-value greater than 0.05.



```
   Dep. Variable:        tuactdur24          R-squared:       0.041
           Model:               OLS      Adj. R-squared:      0.040
          Method:     Least Squares          F-statistic:     145.2
            Date:  Sun, 23 Feb 2020  Prob (F-statistic):  8.94e-122
            Time:          12:19:13      Log-Likelihood:    -71869.
No. Observations:             13700                 AIC:  1.437e+05
    Df Residuals:             13695                 BIC:  1.438e+05
        Df Model:                 4
 Covariance Type:         nonrobust

                     coef  std err        t  P>|t|  [0.025   0.975]
      Intercept   98.6877    1.737   56.819  0.000  95.283  102.092
haschild[T.True]  -0.7099    1.002   -0.708  0.479  -2.674    1.255
           teage  -0.5578    0.027  -20.605  0.000  -0.611   -0.505
        tehruslt  -0.1176    0.020   -6.028  0.000  -0.156   -0.079
        tespuhrs  -0.0820    0.018   -4.586  0.000  -0.117   -0.047

      Omnibus:  3163.839      Durbin-Watson:     1.997
Prob(Omnibus):     0.000   Jarque-Bera (JB):  6503.445
         Skew:     1.372           Prob(JB):      0.00
     Kurtosis:     4.967           Cond. No.      268.
```

Conclusion

Through comparison of CDF, hypothesis testing, and simple regression, women's exercise habits are different with those that have children under the age of 18 and those that do not. Those with children exercise less than those that do not have children. However, when looking at multiple regression, other variables account for female exercise habits that do not involve having children.

Resources

DeJesus, John. (2019). What, Why, and How to Read Empirical CDF. *Towards Data Science*. Retrieved from https://towardsdatascience.com/what-why-and-how-to-read-empirical-cdf-123e2b922480

Downey, A. B. (2015). *Think Stats: Exploratory Data Analysis* (2nd ed.). Sebastopol, CA: O'Reilly Media Inc.

Minitab Blog Editor. (2013). How to Interpret Regression Analysis Results: P-values and Coefficients. Retrieved from https://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients

Pathak, Manish. (2018). Joining DataFrames in Pandas. *DataCamp*. Retrieved from https://www.datacamp.com/community/tutorials/joining-dataframes-pandas

U.S. Bureau of Labor Statistics. (2017). American Time Use Survey: Multi-Year Survey Microdata Files from 2003-2015. Retrieved from https://www.kaggle.com/bls/american-time-use-survey/data

VanderPlas, J. (2016). *Python Data Science Handbook: Essential tools for Working with Data (1st ed.).* O'Reilly Media Inc.