Maayan Huber-Karl

https://www.linkedin.com/in/maayan-huber-karl/

Abstract

In 2011 I submitted my thesis on the relations between premature birth and key contributing factors. SPSS was the main tool used for data analysis.   
In 2022 I took a course in modern data analysis which included tools such as Excel, Python, SQL and Power BI. After the course I decided to revisit my thesis with this modern tool set.  
This report explores the use of Excel for running key analyses from my thesis.

V1.1

Data analysis report

Revisiting my thesis: using Excel to analyze contributing factors to premature births

Table of Content

[Introduction 1](#_Toc137327699)

[Data description 2](#_Toc137327700)

[Research population 2](#_Toc137327701)

[Explanatory variables 2](#_Toc137327702)

[Special notes 2](#_Toc137327703)

[Dataset 2](#_Toc137327704)

[Data processing 2](#_Toc137327705)

[Wrangling, cleaning, error correction 2](#_Toc137327706)

[Discarding 3](#_Toc137327707)

[Renaming 3](#_Toc137327708)

[Re-coding 3](#_Toc137327709)

[Categorization 4](#_Toc137327710)

[Data analysis 4](#_Toc137327711)

[Descriptive analysis of the research population 4](#_Toc137327712)

[Comparative analysis of the two research groups 4](#_Toc137327713)

[Multivariate analysis 5](#_Toc137327714)

[Excel file description 5](#_Toc137327715)

[Results 7](#_Toc137327716)

[Univariate Analysis 7](#_Toc137327717)

[Bivariate Analysis 11](#_Toc137327718)

[Discussion of the results 17](#_Toc137327719)

[Univariate Analysis 17](#_Toc137327720)

[Bivariate Analysis 18](#_Toc137327721)

[Impressions from Excel 24](#_Toc137327722)

[Appendix 25](#_Toc137327723)

[Computation of Chi-Square tests 25](#_Toc137327724)

# Introduction

The main research goal in my thesis (data collected 2010-2011) was to analyze the relation between premature (aka preterm) birth and pregnancy monitoring (i.e. prenatal care) as measured either by the trimester monitoring started, or by Kessner index.

We note that at the time of research, ‘trimester monitoring started’ (also named in our research Pregnancy Tracking Start) was not a well-researched factor, with most of the relevant research works dated in 1985 or earlier, with contradicting results.

Additional goals in the research were to analyze and compare the two research groups, preterm birth and normal birth, as well as to analyze the relations between preterm birth and other explanatory variables suggested in the literature such as medical conditions, age and health habits (of the mother).

The methodology of the data collection is described in the thesis and is out of scope for this report.

# Data description

The original data collected as part of the thesis work appears on sheets Data1, Data2 and Data3 of the Excel file accompanying this report.

## Research population

Research population included 312 subjects, of which ~30% are preterm birth based on birth week.

## Explanatory variables

Demographic (age, education, birthplace, religion, religiousness, socioeconomic status), clinical (medical conditions, IVF), behavioral (mother/partner smoking, alcohol consumption), prenatal care (Kessner index, start of prenatal care).

Note: Kessner Index is defined based on combination of values on (1) time of entry into prenatal care (PNC), (2) number of prenatal visits, and (3) gestational age at delivery. More details in the thesis.

## Special notes

Several medical conditions were grouped into a single indicator variable called *chronic diseases*. Further details can be found in the thesis.

## Dataset

The original research data was inserted as-is (with minimal edits) into three Excel sheets (named Data1, Data2, and Data3), instead of being joined (or edited for additional clarity), for backward consistency. All three sheets can be seen as a single table (the split between sheets is column-based). Additional clarity would be provided via the sheets created in this work.

# Data processing

## Wrangling, cleaning, error correction

* Original notations, markings, naming, language and unused variables (as they appear in sheets Data1, Data2, and Data3) were mostly kept unaltered.
* Blanks (i.e. missing values) in the source data (sheets Data1, Data2, Data3) were replaced with “null” value to support PIVOT table analysis.
* On each data sheet (Data1, Data2 and Data3) a line was added at the bottom for counting the number of “null” values on each column.
* Variable *age\_pregnant\_week* (i.e. birth week) – a value was corrected on row 167 from “39-1” to “39+1”, interpreted as a typo since only ‘+’ sign is used in indication of birth week.
* Indicator variable *mahalot\_cronicorzihum* (i.e. chronic diseases) on sheet Data1 – a value was corrected on row 107 from 2 (which is out of range) to 1 (which means “yes”). Correction was based on value of variable mahalot\_name (i.e. disease name) for same subject which was 3 (as opposed to no disease which would have a null value).
* Variable *ishun\_woman* – a value was corrected on row 15 from 3 (an illegal value) to “3b” (which means smoking less than before the pregnancy) based on a non-null value provided under *ishun\_kama\_pahot* (i.e. smoking less by how much).

## Discarding

* Variables with very large percentage of null values were not analyzed in this report (or in the thesis).

## Renaming

Some of the original variable names (on sheets Data1, Data2 and Data3) were transliterations. I decided to keep them in the original sheets (only) for easy comparisons with the original results.

In the new sheets I used proper English names, e.g.:

* *mahalot\_cronicorzihum* 🡪 *Mother\_Chronic\_Or\_Hormonal\_Diseases*
* *dat* 🡪 *Mother\_Religion\_Ethnicity*
* *datiyut* 🡪 *Mother\_Religiousness*
* *kalkali* 🡪 *Mother\_Economic\_Status*
* *ishun\_woman* 🡪 *Mother\_Smoking*
* *ishun\_ben zug* 🡪 *Partner\_Smoking*
* *maakav\_preg\_part* 🡪 *Pregnancy\_Tracking\_Start*

## Re-coding

Several categorical variables were re-coded to replace numeric (or alpha-numeric) values with descriptive strings of text.

Most of the re-coding was done in the ‘Rename and Re-code’ sheet. Since the spreadsheets that contain the raw data do not detail the meanings behind each numeric value, the thesis document was consulted in order to recreate those.

These variables include:

* *Mother\_Birth\_Land*: 1🡪”Israel”, 2🡪”the former Soviet Union”, 3🡪”Else”
* *Mother\_Chronic\_Or\_Hormonal\_Diseases*: 0🡪”No”, 1🡪”Yes”
* *Mother\_Religion\_Ethnicity:*  1🡪"Muslim Arab", 2🡪"Christian Arab", 3🡪"Russian Christian", 4🡪"Circassian", 5🡪"Druze", 6🡪"Ethiopian", 7🡪"Ashkenazi", 8🡪"Mizrachi", 9🡪"Other"
* *Mother\_Religiousness:*  1🡪"Religious", 2🡪"Traditional", 3🡪"Secular"
* *Mother\_Economic\_Status:* 1🡪"Below average", 2🡪"Average", 3🡪"Above average"
* *Mother\_Smoking:*  1🡪"Never smoked", 2🡪"Stopped smoking completely", "3b"🡪"Smokes less than before pregnancy", "3a"🡪"Smokes today as before pregnancy"
* *Partner\_Smoking:* 0🡪"Doesnt smoke", 1🡪"Smoke"
* *Pregnancy\_Tracking\_Start*: 1🡪"First trimester", 2🡪"Second trimester", 3🡪"Third trimester"
* *Mother\_Education*: 1🡪"Elementary", 2🡪"High-school", 3🡪"Academic"
* *Kessner\_Index*: 1🡪"Adequate pregnancy monitoring", 2🡪"Medium pregnancy monitoring", 3🡪"Inadequate pregnancy monitoring"
* *Pre-Pregnancy Consultation*: 0🡪”No”, 1🡪”Yes”
* *In\_Vitro\_Fertilization*: 0🡪”No”, 1🡪”Yes”

## Categorization

The following categorization was performed (sheet ‘Categorize’):

* Categorical variable *Mother\_Age\_Category* was created based on the mother’s age (at the time of birth) divided into three ranges: <20, 20<= and <40, >=40. The following category names were selected respectively “<20”, “20-39”, “>=40”.
* Categorical variable *weight\_category* (of the baby at the time of birth) was created based on three ranges: <2.5kg, 2.5<= and <=4kg, >4kg. The following category names were selected respectively “Low”, “Normal”, “High”.
* Categorical variable *prematurity\_by\_weight\_baby* was created based on the weight at the time of birth divided into two ranges: <=2.5kg, >2.5kg. The following category names were selected respectively “prematurity”, “not\_prematurity”.
* Categorical variable *prematurity\_by\_birth\_week* (i.e. indicator of preterm/normal birth) was created based on two ranges: before week 37, week 37 and after. The following category names were selected respectively “premature”, “not premature”.

General notes:

* The ranges selected are based on relevant literature (see Thesis). Note this includes the difference in boundary settings e.g. between age and weight (>= vs > for the upper category).
* For the rest of the analysis, unless specified otherwise, prematurity of birth is determined based on the birth week.

# Data analysis

## Descriptive analysis of the research population

Average, min, max, and standard deviation were computed for numeric variables (sheets ‘Weight\_baby’ and ‘Mother\_age’). Histograms of their distribution were also computed and displayed in those sheets.

Pivot tables and pie charts were computed for categorical variables (sheet ‘PIVOTs’)

## Comparative analysis of the two research groups

First, descriptive analysis was performed on each group (preterm and normal) separately for categorical (and categorized) variables (sheet ‘Pivot by Prematurity) and for numeric variables (sheets ‘Welch and Histogram weight\_baby’ and ‘Welch and Histogram mother age’).

Next, we analyzed the effects of explanatory variables on preterm birth. Tests of significance used were chi-square test of homogeneity for categorical variables (sheet ‘Chi-Square’), two-tailed Welch's t-test (difference of means, assumed unequal variances) for numeric variables (sheet ‘Welch and Histogram mother age’). Confidence Level of 95% was used.

As a sanity check on our data, since prematurity was determined by week of birth, we wanted to verify that baby weight was significantly different between the two research groups. We verified it using both Welch’s t-test for the continuous variable (sheet ‘Welch and Histogram baby weight’) and Chi-square test for the categorized variable (sheet ‘Chi-Square prematurity’).

## Multivariate analysis

The thesis includes logistic regressions computed for select explanatory variables. These are out of scope for this report.

# Excel file description

Sheets:

* *Data1*, *Data2*, *Data3* - raw dataset (with minimal alterations as described above).
* *Categorize* – This is where we defined new categorical variables based on continuous variables. See more details under Categorization.
* Note in the computation of *prematurity\_by\_birth\_week* we transformed text (e.g. “34+6”) into numeric count of days (244 in this example) and compared to week 37 represented in days (i.e. 259).
* *Rename and Re-code*
  + In this sheet, we perform renaming and re-coding of variables as detailed in the Renaming and Re-coding sections.
* *Weight\_baby* –
  + Numeric variable *weight\_baby* was divided into bins of 100gr to compute a histogram. Null values were discarded.
  + Frequencies were computed using Excel’s “Histogram” method (under “Data Analysis”), part of the “Analysis ToolPak” addon.
  + Histogram was visualized as a line graph called “Frequency Weight\_Baby”.
  + A second graph called “Probability Weight\_Baby” displays the approximate Normal distribution of the binned weight data. Excel function NORM.DIST was used.
  + Basic statistics (Min, Max, Average, StdDev) were calculated.
* *Age Hist* –
  + Numeric variable *Mother\_Age* was already discrete (recorded as integer), so it fit naturally into bins of 1 year to compute a histogram. Null values were discarded.
  + Histogram was displayed as a line graph called “Frequency Mother\_Age”.
  + A second graph called “Probability Mother\_Age” displays the approximate Normal distribution of the binned age data. Excel function NORM.DIST was used.
  + Basic statistics (Min, Max, Average, StdDev) were calculated.
* *PIVOTs* –
  + This sheet presents the results of the descriptive analysis of variables as detailed in the “Descriptive analysis of the research population” section.
  + Pivots tables were used to present count and percentage of each category for categorical variables.
    - Categorical variables analyzed in the Pivot tables: Mother\_Age\_Category, Mother\_Chronic\_Or\_Hormonal\_Diseases, Mother\_Birth\_Land, Mother\_Smoking, Partner\_Smoking, Mother\_Religion\_Ethnicity, Mother\_Religiousness, Mother\_Education, Mother\_Economic\_Status, Pregnancy\_Tracking\_Start, Kessner\_Index, Pre-Pregnancy Consultation, In\_Vitro\_Fertilization.
  + Pie charts were created for all variables.
* *Pivot by Prematurity* –
  + This sheet displays side-by-side the distributions (in the forms of Pivot table and bar chart) of select categorical variables, as detailed in the ‘Data analysis’ section under ‘Comparative analysis of the two research groups’.
  + Variables analyzed – Prematurity by Mother Age, Prematurity by Mother Birth Land, Prematurity by Mother Chronic Or Hormonal Diseases, Prematurity by Mother Religion Ethnicity, Prematurity by Mother Religiousness, Prematurity by Mother Education, Prematurity by Mothe Economic Status, Prematurity by Mother Smoking, Prematurity by Partner Smoking, Prematurity by Pregnancy Tracking Start, Prematurity by Kessner, Prematurity by Pre-Pregnancy Consultation and Prematurity by In Vitro Fertilization.
* *Chi-Square prematurity* –
  + This sheet details the “sanity check” analysis performed on the weight\_category variable, as detailed in ‘Comparative analysis of the two research groups’.
  + We used the function CHISQ.TEST() to compute the chi-square p-value.
  + weight\_category - this variable was created in the ‘Categorize’ sheet.
  + prematurity\_by\_birth\_week - this variable was created in the Categorize sheet.
* Observed - a pivot table created to sum up the observed data per weight category for each research group - premature and not premature.
* Expected - expected values. Computed as described in ‘Computation of Chi-Square tests’.
* (obs-exp)^2/exp - a table created to support computation of the chi-square statistic.
* *Chi-Square* –
  + This sheet has been put to “hide”. See “Chi-Square corrected” sheet.
  + This sheet details the chi-square analysis performed on select categorical variables, as detailed in ‘Comparative analysis of the two research groups’.
  + Tables titles “Observed” - pivot tables created to sum up the observed data per category for each research group - premature and not premature.
  + Tables titled “Expected” - expected values. Computed as described in ‘Computation of Chi-Square tests’.
  + Tables titled “(obs-exp)^2/exp” - tables created to support computations of the chi-square statistic.
  + We used the function CHISQ.TEST() to compute the chi-square p-value. Values under 0.05 were highlighted in green indicating significant difference between observed and expected values, meaning there’s a dependency between the dependent and the explanatory variables. Values over 0.05 were highlighted in red.
  + The following variables were analyzed: *Partner\_Smoking, Pregnancy\_Tracking\_Start, In\_Vitro\_Fertilization, Mother\_Age\_Category, Mother\_Birth\_Land, Mother\_Religion\_Ethnicity, Mother\_Religiousness, Mother\_Education, Mother Economic status, Mother\_Smoking, Kessner\_Index*, IN-Vitro Fertilization, Pre-Pregnancy Consultation.
  + Variables Pregnancy\_Tracking\_Start and Kessner\_Index are highlighted because they were the focus of the research.
  + Important notes:
    - Null values were filtered out of the pivot tables so computed sum values for columns and rows (“Grand Total”) disregard them (important for computing expected values). Null values are not considered a category for analysis but represent missing value.
    - Pivot tables show empty values instead of 0 by default. It’s important for the correct computation of CHISQ.INV.RT() and CHISQ.TEST() functions to not have missing values in the observed values (i.e. Pivot tables). This can be corrected from the Pivot table options by setting 0 as value for “For empty cells show”.
* *Chi-Square corrected* –
  + Same as Chi-Square but with a couple of corrections.
  + Variable Pre-Pregnancy Consultation was discarded as it is degenerated. Technically a binary variable, one of its categories includes only 3 subjects, disqualifying it from a reliable Chi-Square analysis.
  + Variable Mother Religion Ethnicity – categories with 3 or less total subjects were removed from the analysis, correcting the result from significant to non-significant.
* Welch and Histogram weight\_baby –
  + Variable weight\_baby was divided between research groups – preterm and normal birth, based on the week of birth.
  + To determine the difference of means between the two groups, we applied a t-test as detailed in ‘Comparative analysis of the two research groups’ in the Results chapter. Recall this was a sanity test to verify a certain level of consistency between the two different definitions of a preterm birth.
  + We also computed and plotted the histogram of each group. The graph is also show in ‘Comparative analysis of the two research groups’ as part of the t-test results.
  + In this sheet we observed a large difference of variances between the two groups, which matched our assumptions (or rather non-assumption of equal variances) for applying a (two-tailed) Welch’s t-test.
* Welch and Histogram mother\_age –
  + Variable Mother\_Age was divided between research groups - premature (before week 37) and not premature (week 37 and after).
  + To determine the difference of means between the two groups, we applied a t-test as detailed in ‘Comparative analysis of the two research groups’ in the Results chapter.
  + We also computed and plotted the histogram of each group. The graph is also show in ‘Comparative analysis of the two research groups’ as part of the t-test results.
  + In this sheet we observed a small difference of variances between the two groups, yet we still applied a (two-tailed) Welch’s t-test because we did not assume equal variances.

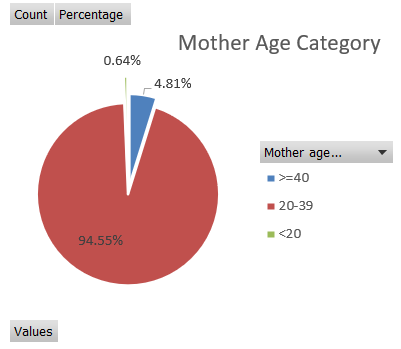
Excel functionality

* To compute histogram for the numeric variable weight\_baby, the Analysis ToolPak add-in was loaded as described here: <https://www.excel-easy.com/examples/histogram.html>
* To compute ‘t-Test: Two-Sample Assuming Unequal Variances’, the Analysis ToolPak add-in was used.
* NORM.DIST function was used to generate an approximate Normal distribution of histograms (sheets ‘Weight Hist’, ‘Age Hist’).
* CHISQ.TEST function was used to perform a chi-squared test of homogeneity and compute its p-value (sheets ‘Chi-Square prematurity’ and ‘Chi-Square’)

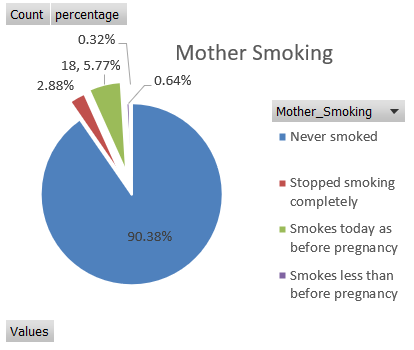
# Results

## Univariate Analysis

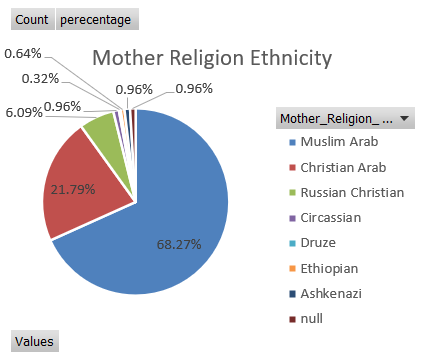
Categorical Variables – below are the pie charts visualizing the distribution of values of categorical variables. Preliminary notes and observations are followed.

A screenshot of a graph

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer screen

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidenceA screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

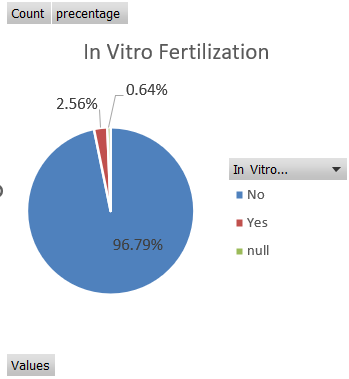
Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with low confidenceA screenshot of a graph

Description automatically generated with low confidence

A screenshot of a graph

Description automatically generated with medium confidence 

A picture containing text, screenshot, diagram, font

Description automatically generated

A picture containing text, screenshot, font, diagram

Description automatically generatedA picture containing text, screenshot, font, logo

Description automatically generated

Preliminary notes and observations:

* Mother Age Category – population is mostly young (under 40).
* Mother Chronic Diseases – population is mostly with no background diseases.
* Mother Smoking – most of population never smoked.
* Partner Smoking – divides largely half and half.
* (Pie Chart not included) Mother drinking alcohol – no alcohol drinking was reported. This does not allow us to run any statistical test about the contribution of this factor to predicting preterm birth. Further discussion can be found in the Thesis.
* Mother Religion / Ethnicity – most of the population is either Christian or Muslim Arab.
* Mother Economic Status – almost half did not answer. Most of the subjects who did, are either average or below.
* Kessner Index – most of the population’s pregnancy tracking was not considered adequate.
* Pre-Pregnancy Consultation – most of the population did not receive pre-pregnancy consultation. Also, there’s a large percentage (~17%) of null values, i.e. unreported value.
* In Vitro Fertilization – most of the pregnancies were not IVF.
* Baby Weight – most of the population had normal or low baby weight (i.e. below 4.0 kg).

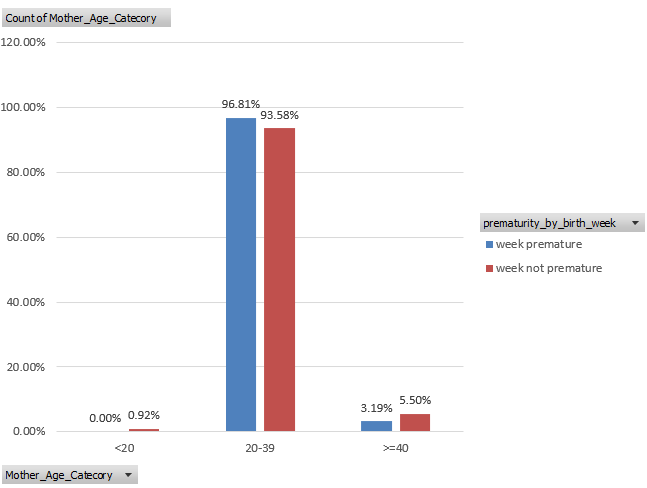
Numeric Variables:

* Weight hist –
  + Based on the histogram graph shape, the data roughly follows a bell-shaped distribution with a left tail representing preterm births.
  + The histogram is noisy (not smooth) compared to the ideal bell-shaped normal distribution / graph.
  + The Thesis discusses in more detail the question of is the research data representative, highlighting the fact that the research population was sampled entirely from a single hospital.
* Age hist
  + The histogram shows that the age distribution is largely bell-shaped and slightly skewed to the right (i.e. right tail representing older mothers) as expected.
  + Note that based on the definition of age categories, most mothers are still considered either young or normal (i.e. being in the two lower of the three age categories, i.e. below 40).

## Bivariate Analysis

Bar charts

These charts provide visual comparison of the two research groups (preterm and normal birth) per explanatory categorical variable. The percentages are taken as total out of each research group separately, e.g. for Mother Age Category, 96.81% **of preterm births** are between ages 20-39.

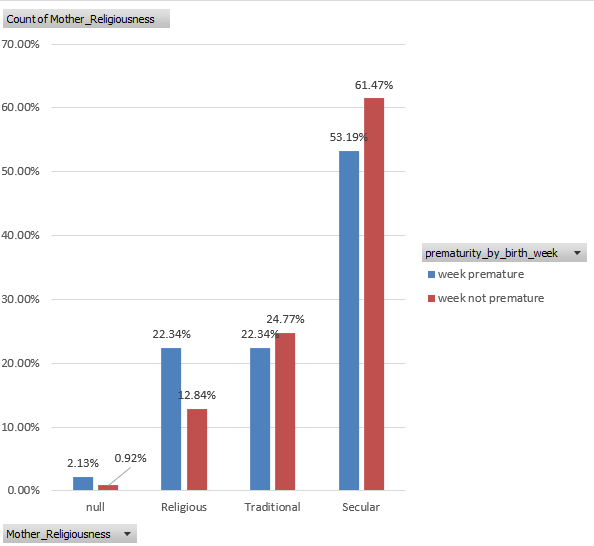
 A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a computer screen

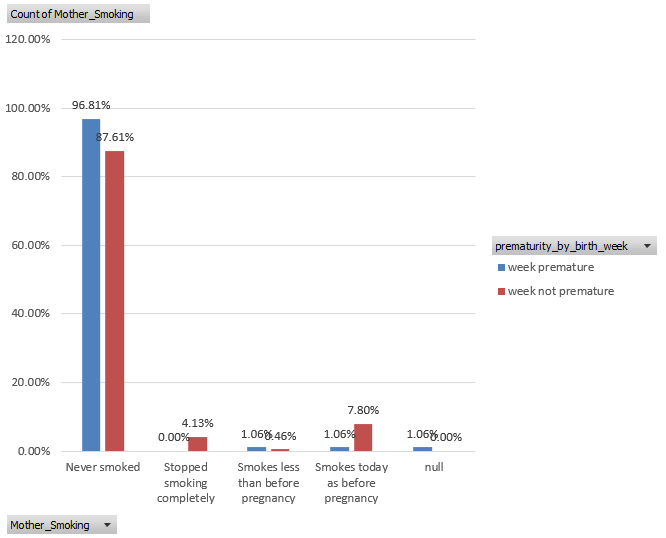
Description automatically generated with low confidence A screenshot of a computer screen

Description automatically generated with low confidence

 A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

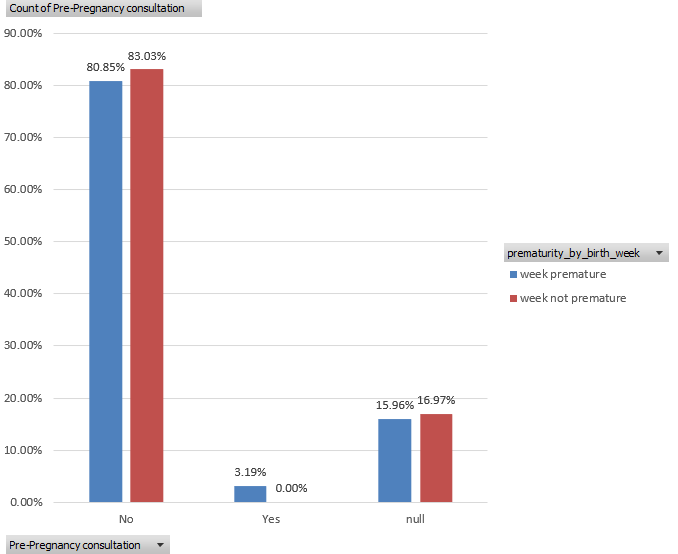
Description automatically generated with medium confidence 

A screenshot of a graph

Description automatically generated with medium confidence A screenshot of a graph

Description automatically generated with medium confidence

A screenshot of a graph

Description automatically generated with medium confidence 

A screenshot of a graph

Description automatically generated with medium confidence

Preliminary notes and observations:

* Mother Age Category – no noticeable differences in distribution
* Mother Birth Land – no noticeable differences in distribution
* Mother Chronic Diseases – noticeable differences in distribution, with preterm birth roughly x4 more likely for mothers with chronic diseases.
* Mother Religion / Ethnicity – no noticeable differences in distribution, perhaps except for Russian Christian population being ~7% of preterm births vs being ~3% of normal births.
* Mother Religiousness – noticeable differences in distribution, with preterm birth ~74% more likely for religious mothers.
* Mother Education - noticeable differences in distribution, with the largest difference in favor of preterm births observed for academic population.
* Mother Economic Status – very large percentage (more than 40%) of unreported values.
* Mother Smoking – noticeable differences in distribution, with preterm birth roughly x8 less likely for mothers who did not change their smoking habits and no preterm births observed for mothers who stopped smoking when they became pregnant.
* Partner Smoking – noticeable differences in distribution, with preterm birth more likely for mothers with non-smoking partner and less likely for smoking partners.
* Pregnancy tracking start – noticeable differences in distribution, with preterm birth least likely (relatively) for early tracking, most likely (relatively) for tracking starting second trimester, and no major differences for late tracking.
* Mother drinking alcohol – noticeable differences in distribution, with preterm birth less likely (relatively) for medium-adequacy tracking compared with non-adequate tracking. Adequate tracking almost non-observed within the research population.
* Kessner Index – we observe that inadequate monitoring is correlated with higher rates of preterm births.
* Pre-Pregnancy Consultation – a noticeable percentage of null (unreported) values (~16%) and a very low number of subjects (3 in total) who received pre-pregnancy consultation, effectively making this a single-category variable.
* In Vitro Fertilization – noticeable differences in distribution, with preterm birth roughly x16 (!!) more likely for mothers who had IVF.

Chi-Square tests:

Recall the significance level was set at 0.05.

[Significant] For the following variables the p-value was smaller than the significance level, meaning we observed a significant difference in the values of the explanatory variables between the two research groups (preterm and normal birth), suggesting a relationship with the dependent variable (prematurity by birth week) –

* Clinical variables –
  + Mother\_Chronic\_Or\_Hormonal\_Diseases (0.044)
  + Partner\_Smoking (0.011)
  + In\_Vitro\_Fertilization (0.0003)
  + Mother\_Smoking (0.018)
* Pregnancy monitoring variables –
  + Pregnancy\_Tracking\_Start (0.019)
  + Kessner\_Index (0.044)

[Non-significant] For the following variables the p-value was larger than the significance level –

* Demographic variables –
  + Mother\_Birth\_Land (0.89)
  + Mother\_Religiousness (0.096)
  + Mother\_Education (0.069) - **Note it is “not too far” to the significance level**
  + Mother\_Economic\_Status (0.20)
  + Mother\_Religion\_Ethnicity (before correction - 0.036, after correction – 0.39) – for this variable we corrected the Chi-Square analysis by discarding categories with 3 or less total observations (these are Ashkenazi, Circassian, Druze, Ethiopian).
* Clinical variables –
  + Mother\_Age\_Category (0.435)

[Non-significant] For the following variables we did not have enough data to run an analysis –

* Pregnancy monitoring variables –
  + Pre-Pregnancy Consultation (before correction - 0.008, after correction – N/A) – since only 3 subjects received pre-pregnancy consultation, we did not have enough data to run a Chi-Square analysis.

T-tests (Welch’s) and Histogram:

Mother Age

* A Welch’s t-test was performed to analyze the relationship of explanatory continuous variables (mother age) and the dependent variable (prematurity by birth week).
* With significance level set at 0.05, the results for mother age variable (t = -0.1031, p = 0.9179) imply there’s no significant difference of means, meaning no relationship.
* Histogram –
  + The histogram shows that the age distribution is largely bell-shaped and slightly skewed to the right (i.e. right tail representing older mothers) for both study groups, as expected.
  + Note that based on the definition of age categories, most mothers are still considered either young or normal (i.e. being in the two lower of the three age categories, i.e. below 40).
  + We also see that the two histograms largely overlap, providing a clear visual that aligns with the t-test analysis, that indeed there’s no significant difference of means (and in fact no significant difference overall) between the two groups.

Baby Weight

* A Welch’s t-test was performed to analyze the relationship between the continuous variables baby weight and prematurity by birth week.
* Note that Baby Weight is NOT an independent variable of preterm birth, and in fact highly correlated. This test can be regarded as a sort of sanity check.
* With significance level set at 0.05, the results for baby weight variable (t = -16.728, p = 1.61272951414871E-34) imply there is a significant difference of means, meaning there’s a relationship, as expected.
* Histogram –
  + We see significant differences between the distributions of the two research groups. While both are largely bell-shaped, the preterm births have lower mean and larger variance, while the normal births have a higher mean and a smaller variance.
  + The histogram (i.e. distribution of recorded values) is noisy compared with the ideal bell-shaped normal distribution.
  + Notice there’s an overlap between the two groups. It’s possible for a preterm birth (defined by the week of birth) to have a high birth weight, and for a normal birth to have a low birth weight.

# Discussion of the results

## Univariate Analysis

Preliminary / trivial discussion of results observed during Univariate Analysis can be found in the Results chapter.

Additional notes:

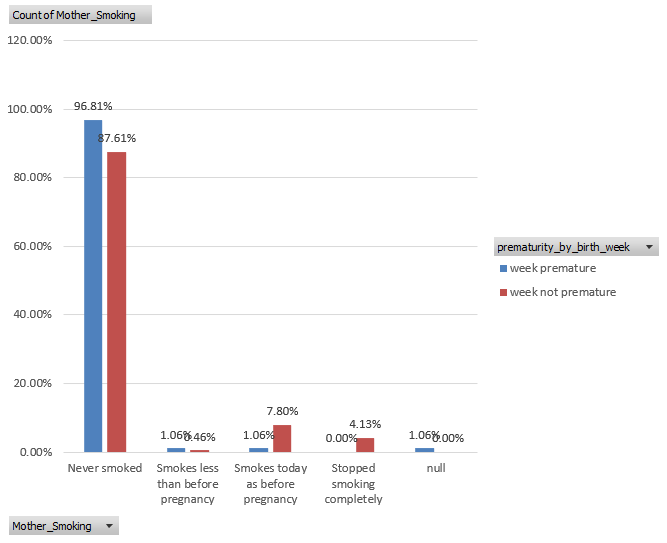
* Mother Age Category – population is mostly young (under 40). The global population stats were not included in the original report for comparison. Also, it would be interesting to analyze potential changes in these stats since the time of research as fertility treatments (including IVFs) progress and become more accessible over time, potentially increasing the chances of successful pregnancies in older age.
* Mother Smoking – most of population (~90%) reported as “never smoked”. This stat should be treated with skepticism, as it relied only on self-reporting with no independent validation. Moreover, the data collection was specifically for cigarette smoking and did not include (although it may have needed to) shisha smoking.
* Mother Religion / Ethnicity – most of the population is either Christian or Muslim Arab. This is uncharacteristic of the general population in Israel, but it is characteristic of the specific hospital from which the data was collected.
* Mother Economic Status – almost half did not answer. Most of the subjects who did, are either average or below. This is uncharacteristic of the general population in Israel, but it is characteristic of the specific hospital from which the data was collected.

## Bivariate Analysis

Categorical variables (see p-values in ’Bivariate Analysis’):

Significant –

* Clinical variables –
  + Mother\_Smoking –
    - The graph comparison shows difference between the two research groups, with preterm birth roughly x8 less likely for mothers who did not change their smoking habits (i.e. kept smoking) and no preterm births observed for mothers who stopped smoking when they became pregnant – both of these results make sense and are consistent with literature.
    - We remind that this variable includes cigarette smoking only (does not include shisha smoking).



* + Mother\_Chronic\_Or\_Hormonal\_Diseases
    - We note that in the bar chart we observed noticeable differences in distribution, with preterm birth roughly x4 more likely (than normal birth) for mothers with chronic diseases.
    - We also note the relatively low percentages in that category (representing a total of 8 subjects), which probably explain why the significance was not too far below the significance level. Perhaps if the research included more subjects with chronic diseases, we could have observed this dependency more significantly.

A screenshot of a computer screen

Description automatically generated with low confidence

* + Partner\_Smoking
    - We note that while this variable was found to be significant, the results are opposite to expected, with smoking partner being correlated with a smaller chance of preterm birth.
    - Most likely explanation comes after examining the questionnaire and the raw data. In the questionnaire, smoking partner was coded ‘1’ and non-smoking ‘2’, while in the raw data (in the Excel spreadsheet), a non-smoking partner was coded ‘0’ and smoking partner as ‘1’, suggesting an accidental reversal of the coding.

A screenshot of a graph

Description automatically generated with medium confidence

* + In\_Vitro\_Fertilization
    - Results of the Chi-Square analysis match the bar chart, with preterm birth roughly x16 (!!) more likely than normal birth (7 cases compared with 1) for mothers who had IVF.
    - Significant dependency is also observed in the relevant literature.

A screenshot of a graph

Description automatically generated with medium confidence

* + These results match our expectations and the literature.
* Pregnancy monitoring variables –
  + Kessner\_Index -
    - In the bar chart we see that inadequate monitoring is correlated with higher rates of preterm births, as is expected. This is also consistent with the Chi-Square analysis result.

A screenshot of a graph

Description automatically generated with medium confidence

* + Pregnancy\_Tracking\_Start –
    - This was one of the two main variables in focus of our research.
    - The results of the Chi-Square analysis match our expectations. Recall our notes from ‘Introduction’ that this was not a well-researched factor and most of the previous research noted did not find it to be a significant contributing factor.
    - When breaking down the results per trimester, we also see that the chance of preterm birth is the same between preterm and normal birth when pregnancy tracking started in the third trimester. This could potentially be explained if the research population did not include any pregnancy complications that may have led to early birth hadn’t they been detected in a late tracking. In other words, for this research population, late tracking may have been “too late” to have a significant effect on the expected birth week.

A screenshot of a graph

Description automatically generated with medium confidence

Non-significant –

* Demographic variables –
  + Mother\_Religion\_Ethnicity –
    - We observe that the research population is largely homogenic with ~90% Arab ethnicity subjects on both research groups, with religion (Christian vs Muslim) apparently not affecting the preterm birth rates inside that ethnic group.
    - Most of the other groups are rare (3 or less total observations), and thus have been discarded to not invalidate the Chi-Square analysis (observed x2 to x4 relative proportions lead to falsely computed significant relationship).
    - Only other group kept in the analysis was Russian Christian.
    - Relevant literature at the time of research also does not demonstrate ethno-religious association as a significant contributing factor.

A screenshot of a computer screen

Description automatically generated with low confidence

* + Mother Birth Land – was not found to be significant in our research, as is the case in literature at the time of research.
  + Mother\_Religiousness
    - While the Chi-Square computation did not show significant differences (as is the literature at the time of research), we do observe a trend, that is that more secular population tends to have lower chance of preterm birth and vice versa. Note the relative difference is largest for religious mothers, with preterm birth 74% more likely than normal birth.

A screenshot of a computer screen

Description automatically generated with low confidence

* + Mother\_Education –
    - While this was found as non-significant based on Chi-Square analysis, the p-value was not too far from the significance level, which aligns with the noticeable differences observed in the distribution bars.
    - These results contradict our expectations (and literature) with the largest difference in favor of preterm births observed for academic population.
    - One potential explanation may be a difference in versions of the questionnaire, with earlier version potentially having opposite coding for elementary and academic education.

A screenshot of a graph

Description automatically generated with medium confidence

* + Mother\_Economic\_Status
    - We note that the largest difference (x2) was observed for (above average income subjects) very low percentage values as well as a very small total number of observations, which can explain why it was not enough to establish a significant relationship.
    - We also observe counter-intuitive results for the average and below-average incomes, where the average income subjects had a noticeable higher chance of preterm births, and the below average income subjects had a noticeable lower chance of preterm birth.
    - Most likely explanation might be that almost half of the research population did not report its economic status, and had they reported it, then the results would probably have been aligned with our expectations and the literature.

A screenshot of a graph

Description automatically generated with medium confidence

* Clinical variables –
  + Mother\_Age\_Category – the age categories that were used, although based on common literature practice, did not segment the research population very well (~95% concentrated in the same age category), thus may not have allowed a relationship to be observed (if there was one).

Numeric variables

* The only numeric explanatory variable analyzed was mother age, and was found to have non-significant relationship with target variable.
* This is consistent with the results for the categorized Mother Age variable, which was also not significant.
* We see that most subjects are within the normal (or young) age range (as defined in literature being <40), where no significant negative effects on preterm birth are expected (based on literature).
* Perhaps if older mothers (>=40) had larger representation in the research population, we could have observed a difference there.

Reader can refer to the thesis for further discussion of the above results.

# Impressions from Excel

When considering the use of Excel, the following main advantages and shortcomings come to mind.

Advantages to using excel:

* Excel’s biggest plus is having the data and the analysis in the same place (opposed to manually exporting and importing between different tools, e.g. Excel and SPSS) which saves time, minimizes human errors, makes tracking of computations easier and maintains consistency.
* Excel sheets supports additional information alongside the data table, such as notes and comments, which eases data management.

Downsides to using Excel:

* Having the data, intermediate computations and final results under all in the same file or even sheet, also has a downside i.e. it’s hard to create a clean and well-structured report.
* Excel sheets can host multiple tables and charts/graphs but their layout requires manual effort and the results are not visually appealing or easy to review.
* PIVOT tables require manual refresh to update, which may lead to breaks when renaming variables or their values. Referencing PIVOT tables is also cumbersome and may require manual editing under similar circumstances.
* Not every computation is traceable and amendable, e.g. when computing histogram or t-test via the Analysis ToolPak.
* Excel doesn’t automatically identify and alert on changes in variable name (e.g. ‘\_’ replaced with a space, or lower/upper case switches)
* No support for repetition, e.g. we cannot generate a series of PIVOT tables and charts based on a series of variables. Instead we need work manually.

# Appendix

## Computation of Chi-Square tests

* We compute the Expected tables using this formula:



* To compute the ꭕ2-statistic we compute another table using this formula (then sum):

Text

Description automatically generated

Letter

Description automatically generated