BUSINESS ANALYTICS
ADD HEALTH DATASET DATA MINING

Waterford Institute of Technology

Tsvetoslav Dimov

20077038

BSc (Hons) in Software Systems Development

Year 4

Lecturer: Brenda Mullally

Date: 08/03/2019

Contents

Introduction

This is a report for the module Business Analytics II. It contains information obtained through research and data mining. The dataset being mined is called *The National Longitudinal Study of Adolescent to Adult Health* (Add Health). A research question has been identified. Based on the research question Data Analysis has been carried out and the findings are presented in a mix of univariate and bivariate charts. Lastly, hypothesis tests have been performed to proof given hypothesis true or wrong.

Research Question

One of the first requirements of this project was to pick a dataset. My choice was *The National Longitudinal Study of Adolescent to Adult Health* (Add Health). It is a longitudinal study of a nationally representative sample of adolescents in grades 7-12 in the United States during the 1994-1995 school year. The dataset combines data to do with respondents' social, economic, psychological and physical well-being with conceptual data on the family, neighbourhood, community, school, friendships and more (cpc.unc.edu, n.d.).

The next logical step was to propose a research question and get approval from my lecturer. The research question is the following:

WHAT IMPACT DO PARENTS HAVE ON THEIR CHILDREN'S SMOKING HABITS?

Having picked a dataset and a research question in mind, I began familiarising with the dataset. I inspected it for variables that could potentially be useful for conducting analysis with the aim to answer the research question. The selected variables span across multiple sections:

Section 14: Resident Mother

- H1RM1 - How far in school did she go?

Section 15: Resident Father

- H1RF1 - How far in school did he go?

Section 16: Relations with Parents

- H1WP1 - Do your parents let you make your own decisions about the time you must be home on weekend nights?
- H1WP2 - Do your parents let you make your own decisions about the people you hang around with?
- H1WP3 - Do your parents let you make your own decisions about what you wear?
- H1WP4 - Do your parents let you make your own decisions about how much television you watch?
- H1WP5 - Do your parents let you make your own decisions about which television programs you watch?
- H1WP6 - Do your parents let you make your own decisions about what time you go to bed on weeknights?
- H1WP7 - Do your parents let you make your own decisions about what you eat?
- H1WP9 - How close do you feel to your mother?
- H1WP10 - How much do you think she cares about you?
- H1WP13 - How close do you feel to your father?
- H1WP14 - How much do you think he cares about you?

Section 28: Tobacco, Alcohol, Drugs

- H1TO1 - Have you ever tried cigarette smoking, even just 1 or 2 puffs?
- H1TO2 - How old were you when you smoked a whole cigarette for the first time?
- H1TO7 - During the past 30 days, on the days you smoked, how many cigarettes did you smoke each day?

Data Analysis

Firstly, I loaded the Add Health dataset. To do that I used *Pandas* library. Then I restricted the dataset to observations that know their biological parents and set the decimal points of *Pandas* to be 3.

```
# Load dataset.
addhealth_data = pd.read_csv('addhealth_pds.csv', low_memory=False)

# Restrict dataset to observations that know their biological parents.
dataset = addhealth_data[(addhealth_data['H1NF1'] == 7) & (addhealth_data['H1NM1'] == 7)].copy()
pd.set_option("display.precision", 3)  # Set display results to 0 decimal points.
```

The first analysis I did was to find out how big the dataset is:

```
================================================================
Number of observations/rows in the AddHealth dataset:

6504

================================================================
Number of variables/columns in the AddHealth dataset:

2829
```

After I subset the original dataset (6504 rows), my selection was 3412 rows long, meaning that I reduced the original size by nearly 50%.

Data cleaning is a very important stage of data analysis. If one omits or does not pay enough attention to the details in it, conducting analysis and delivering the requested outcome would unnecessarily challenge and burden the analyst. Data cleaning is the first step to follow data selection and it includes activities like removing/modifying incorrect, incomplete, duplicated and malformatted data.

After observing results from running frequency distributions on a number of variables, I acknowledged that the dataset is almost pristine. The only cleaning activity I had to perform was to replace unnecessary values with null. For example, in variable H1WP2, I replaced the following values with *numpy.nan*, making sure these observations do not count when performing *value_counts()* and other functions by *Pandas:*

2. Do your parents let you make your own decisions about the people you hang around with?			**H1WP2**	num 1
942	0	no		
5420	1	yes		
3	6	refused		
131	7	legitimate skip *[no MOM or DAD]*		
7	8	don't know		
1	9	not applicable		

```
# Section 16: Relations with Parents

# Loop over indices array and replace unnecessary values with null.
for i in [*range(1, 8), 9, 13, 10, 14]:
    dataset[f'H1WP{i}'] = dataset[f'H1WP{i}'].replace([6, 7, 8, 9], numpy.nan)
```

Afterwards, I ran frequency distributions on selected variables, which improved my understanding of the underlying data. An example using variable H1WP9 to show the frequency distributions in percentages. This process was repeated for many other variables:

```
print(
    "\n=========================================
    "How close do you feel to your mother? (%)\n"
    "#1 not at all\n"
    "#2 very little\n"
    "#3 somewhat\n"
    "#4 quite a bit\n"
    "#5 very much\n\n"
    f"{dataset['H1WP9'].value_counts(normalize=True)}")
```

```
=================================================================
How close do you feel to your mother? (%)
#1 not at all
#2 very little
#3 somewhat
#4 quite a bit
#5 very much

5.0     0.686
4.0     0.214
3.0     0.075
2.0     0.023
1.0     0.003
Name: H1WP9, dtype: float64
```

I incorporated extra variables into the dataset. These variables involve the usage of existing variables to calculate and/or categorise data, which in turn extracts more value from the data. They are called derived variables. For example, I used the following question's answers and calculated the number of cigarette packs children smoke per month:

7. During the past 30 days, on the days you smoked, how many cigarettes did you smoke each day?			**H1TO7**	num 2
55	0	no cigarettes		
369	1	one cigarette each day		
249	2	two cigarettes each day		
136	3	three cigarettes each day		
98	4	four cigarettes each day		
138	5	five cigarettes each day		
48	6	six cigarettes each day		
34	7	seven cigarettes each day		

```python
dataset['CIG_MONTHLY'] = dataset['H1TO7'] * 30.42  # Cigarettes per day * average number of days per month.
dataset['CIG_PACKS_MONTHLY'] = round(dataset['CIG_MONTHLY'] / 20)  # Typically a pack contains 20 cigarettes.
dataset['CIG_PACKS_MONTHLY_BINS'] = pd.cut(dataset['CIG_PACKS_MONTHLY'],  # Custom category bins.
                                           [0, 3, 6, 9, 136],
                                           labels=['1-3', '4-6', '7-9', '10+'])
```

```
==========================================================================================
Bins of cigarette packs smoked per month (%):
This is a newly created variable that uses 'pandas.cut()' function to create custom age bins.

1-3    0.431
4-6    0.157
7-9    0.113
10+    0.299
Name: CIG_PACKS_MONTHLY_BINS, dtype: float64
```

From the given percentages one can tell that most of the children either smoke very small number of cigarette packs or a lot them. For this variable, the number of observations is naturally limited to those who smoke. My next step was to find out the mean, standard deviation and other descriptive statistics for the number cigarettes smoked per month:

```
==========================================================================================
Descriptive statistics about number of cigarettes smoked by smokers per month:

count     733.000
mean      194.555
std       248.677
min        30.420
25%        30.420
50%        91.260
75%       304.200
max      2707.380
Name: CIG_MONTHLY, dtype: float64
```

From the descriptive statistics one can tell that there are 733 children that are smokers. The average number of cigarettes they smoke per month is 195. The standard deviation value is close to the mean, which means that most of the data is concentrated in that portion. However, from the max value, one can tell that some students smoke a lot more than the average. This could be due to children not answering this question seriously. The minimum number of cigarettes smoked per month is 30, or 1 per day.

Another new variable that I created by using existing data is called "***PARENT_TYPES***". It determines if children's parents are bossy or soft by collecting answers to questions from H1WP1 through H1WP7. These questions are of format "*Do your parents let you make your own decisions about..*.". The possible answers are two – either yes, or no. If a set of parents let their children make their own decision about 4 or more questions (out of 7), they are soft.

```
# Create a new variable using a subset of the original dataset.
dataset['PARENT_TYPES'] = dataset.loc[:, ['H1WP1', 'H1WP2', 'H1WP3', 'H1WP4', 'H1WP5', 'H1WP6', 'H1WP7']]
    .apply(lambda row: helpers.parents_type(row), axis=1)
```

```
def parents_type(row):
    """
    Determine if the parents are bossy or soft.
    The questions asked are of format: Do your parents let you make your own decisions about...
    Possible answers are:
    0 -- no
    1 -- yes
    A parent is considered soft if they let their child make their own decisions about 4 or more questions
    :param row: Series
    :return: bool
    """
    # Create a dictionary with unique values (1 and 0) and their counts.
    unique, counts = numpy.unique(row.values, return_counts=True)
    counts_dict = dict(zip(unique, counts))
    yes_answers = counts_dict.get(1, 0)  # Get the number of 'yes' answers or replace with 0 if missing.
    return 'Soft' if yes_answers > 4 else 'Bossy'
```

```
====================================
Ratio of bossy to soft parents (%):

Soft        0.692
Bossy       0.308
Name: PARENT_TYPES, dtype: float64
```

From the results one can tell that there are twice more soft parents than there are bossy ones.

Another new variable is called "***PARENTS_CHILD_BOND***". It calculates a bond score, determined by how close children are with their parents and how much they think their parents care about them. The score is of number format and it ranges from 1 (not at all) to 5 (very much). These scores are then binned into 3 categories (Low, Medium and High):

```
# Make a subset that includes ['H1WP9', 'H1WP10', 'H1WP13', 'H1WP14'] columns and apply a function to each row.
dataset['PARENTS_CHILD_BOND'] = dataset.loc[:, ['H1WP9', 'H1WP10', 'H1WP13', 'H1WP14']] \
    .apply(lambda row: helpers.parents_child_bond(row), axis=1)
```

```python
def parents_child_bond(row):
    """
    Calculates how close children are with their parents and how much they think their parents care about them.
    The results of these two calculations is used to determine a bond score.
    Values indicate:
    #1 not at all
    #2 very little
    #3 somewhat
    #4 quite a bit
    #5 very much
    :param row: Series
    :return: numpy.float64
    """
    mother2child = row['H1WP10']
    father2child = row['H1WP14']
    child2mother = row['H1WP9']
    child2father = row['H1WP13']

    # Calculate individual affinities first.
    parents2child = (mother2child + father2child) / 2
    child2parents = (child2mother + child2father) / 2

    return (parents2child + child2parents) / 2
```

```python
# Bin into bonding levels.
dataset['PARENTS_CHILD_BOND_BINS'] = pd.cut(dataset['PARENTS_CHILD_BOND'],
                                            [0, 2, 3.5, 5],
                                            labels=['Low', 'Medium', 'High'])
```

```
=============================================
Levels of bond between parents and their children:

High       0.946
Medium     0.052
Low        0.002
Name: PARENTS_CHILD_BOND_BINS, dtype: float64
```

The results can be interpreted by saying that almost all children have bond with their parents and only a small portion of the children are in bad terms with their parent bodies.

Another new variable is called "***PARENTS_EDU_LEVEL***". It determines the average education level of the parents from answers of the "H1RM1" and "H1RF1" questions for both mother and father respectively. The answer values are sorted by ascending order, which means the higher the answer value is, the higher the education.

1. How far in school did she go?			**H1RM1**	num 2
263	1	eighth grade or less		
568	2	more than eighth grade, but did not graduate from high school		
41	3	went to a business, trade, or vocational school instead of high school		
1811	4	high school graduate		
217	5	completed a GED		
426	6	went to a business, trade, or vocational school after high school		
770	7	went to college, but did not graduate		
1241	8	graduated from a college or university		
512	9	professional training beyond a four-year college or university		
7	10	She never went to school.		

```python
# Make a subset that includes ['H1RF1', 'H1RM1'] columns and apply a function to each row.
dataset['PARENTS_EDU_LEVEL'] = dataset.loc[:, ['H1RF1', 'H1RM1']] \
    .apply(lambda row: helpers.parents_edu_level(row), axis=1)

# Bin into education levels.
dataset['PARENTS_EDU_LEVEL_BINS'] = pd.cut(dataset['PARENTS_EDU_LEVEL'],
                                           [0, 4, 6, 8, 9, 10],
                                           labels=['High-school', 'Vocational', 'Uni', 'Beyond Uni', 'None'])
```

```python
def parents_edu_level(row):
    """
    Determines the average education level of the parents.
    :param row: Series
    :return: numpy.float64
    """
    mother = row['H1RM1']
    father = row['H1RF1']

    return (mother + father) / 2
```

```
==========================================
Parents' level of education:

High-school     1056
Uni              954
Vocational       818
Beyond Uni       406
None               3
Name: PARENTS_EDU_LEVEL_BINS, dtype: int64
```

The results from the frequency distributions can be interpreted by saying that most parent body sets have attended or graduated from high school. The second most common type of education is university. Vocational school have scored closely to university graduates. Only a fraction of the parents has completed a form of education beyond a four-year college degree. The number of parents that have no education is such a small number that it's nearly insignificant.

Visualisation

"Data visualisation is the discipline of trying to understand data by placing it in a visual context so that patterns, trends and correlations that might not otherwise be detected, can be exposed." (medium.com, 2019)

To visualise the variables that I have chosen to analyse I used graphs provided by various Python libraries. They include seaborn, matplotlib and scipy.

Univariate

Creating univariate charts is made simple by using seaborn and matplotlib. However, when creating multiple charts, it can become very repetitive. For this reason, I abstracted the creation of countplots (seaborn graph). This function will be used for every univariate chart:

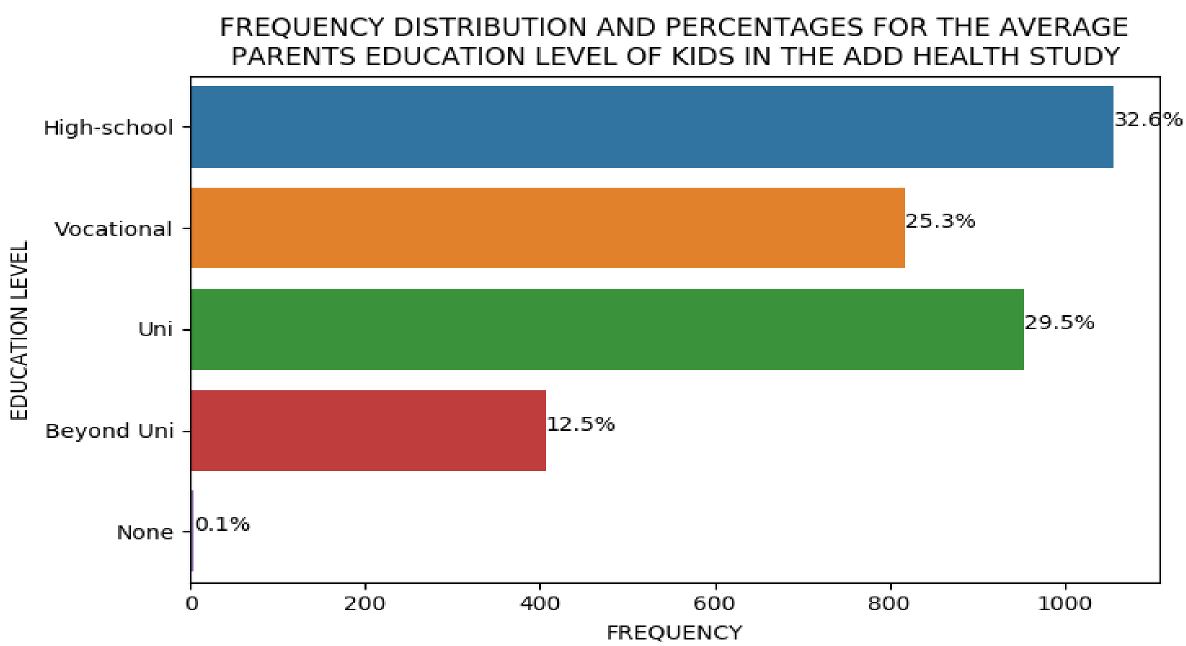
```
def build_countplot(dataset, column_name, title, ylabel, xlabel='FREQUENCY'):
    """
    Abstract the creation and showing of countplot, as it is heavily used
    for my research questions, due to their categorical nature.
    :param dataset: DataFrame
    :param column_name: str
    :param title: str
    :param ylabel: str
    :param xlabel: str
    :return: None
    """
    plt.figure(figsize=(7.5, 4.8))
    ax = seaborn.countplot(y=column_name, data=dataset)
    plt.title(title)
    plt.ylabel(ylabel)
    plt.xlabel(xlabel)
    show_axis_percentages(ax, dataset[column_name])
    plt.show()
```

An additional feature of this plot is the ability to show percentages on the right side of every bar in the chart. This functionality is obtained the following way:

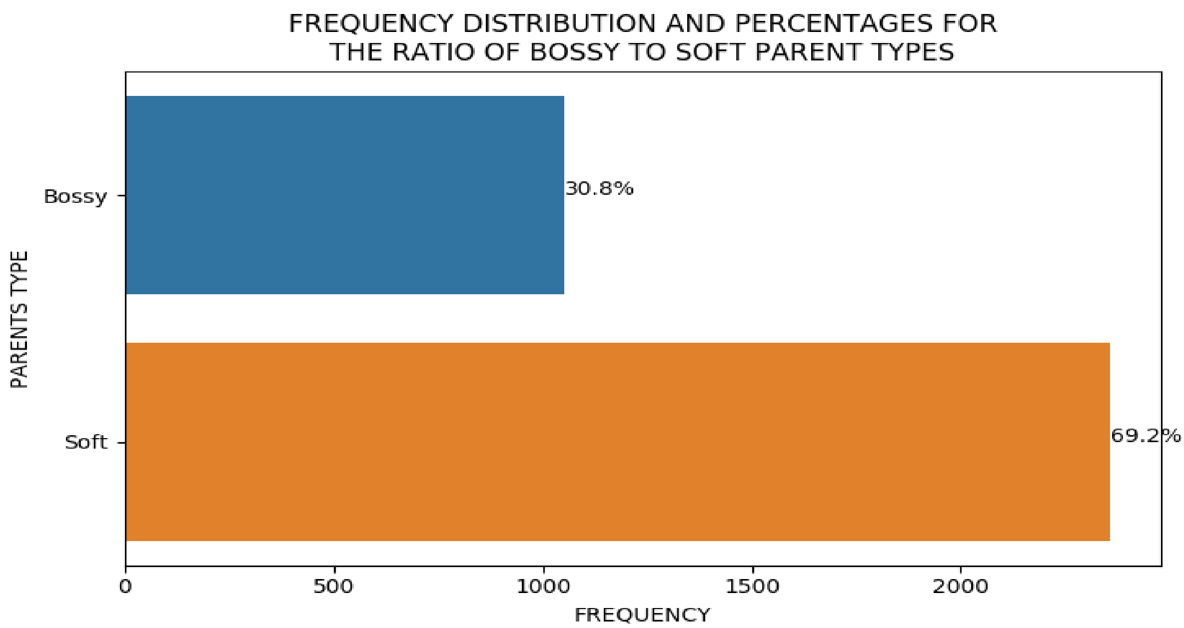
```
def show_axis_percentages(plot, column):
    """
    Helper function that adds percentages to the right of horizontal plot bars.
    :param plot: AxesSubplot
    :param column: Series
    :return: None
    """
    for p in plot.patches:
        percentage = '{:.1f}%'.format(100 * p.get_width() / column.value_counts().sum())
        x = p.get_x() + p.get_width() + 0.02
        y = p.get_y() + p.get_height() / 2
        plot.annotate(percentage, (x, y))
```

The ***build_countplot*** function takes in a DataFrame object, which it will use to retrieve the target column. It also requires a main title and both x and y labels. As seen above, the x label has a default value of "FREQUENCY" as most of the time the graphs will display a frequency on its x-axis. Here is an example usage of the function:

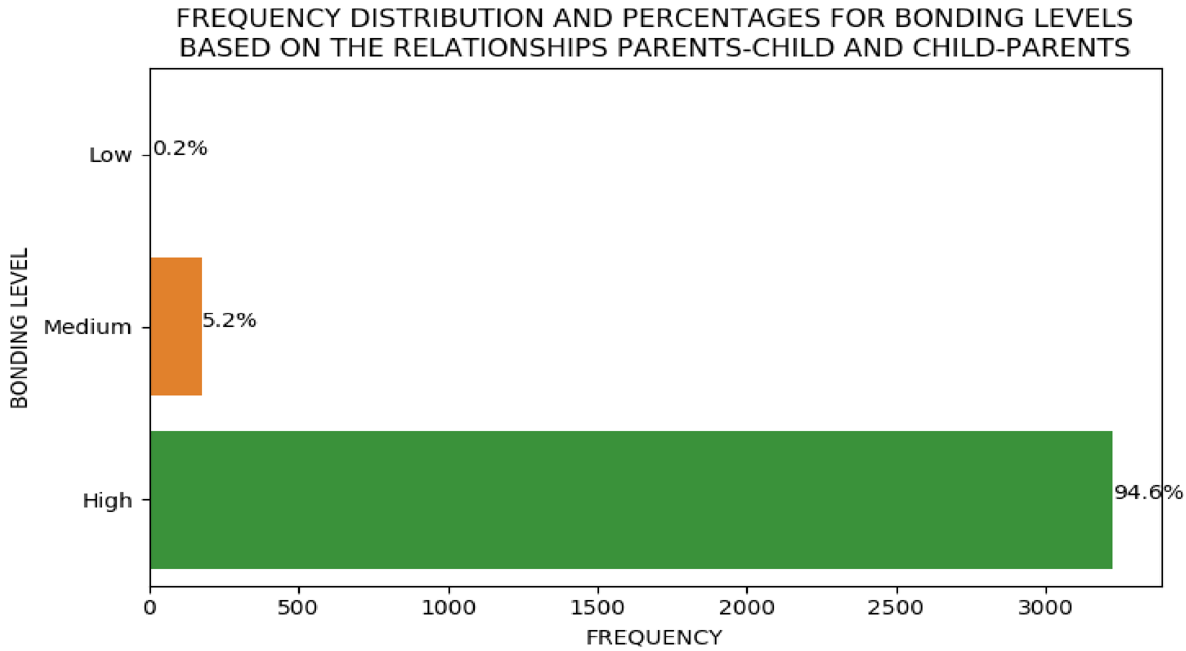
```
# Display a chart showcasing the frequency distribution and percentages
# of parents' education level.
helpers.build_countplot(dataset=dataset,
                        column_name='PARENTS_EDU_LEVEL_BINS',
                        title='FREQUENCY DISTRIBUTION AND PERCENTAGES FOR THE AVERAGE\n'
                              'PARENTS EDUCATION LEVEL OF KIDS IN THE ADD HEALTH STUDY',
                        ylabel='EDUCATION LEVEL')
```



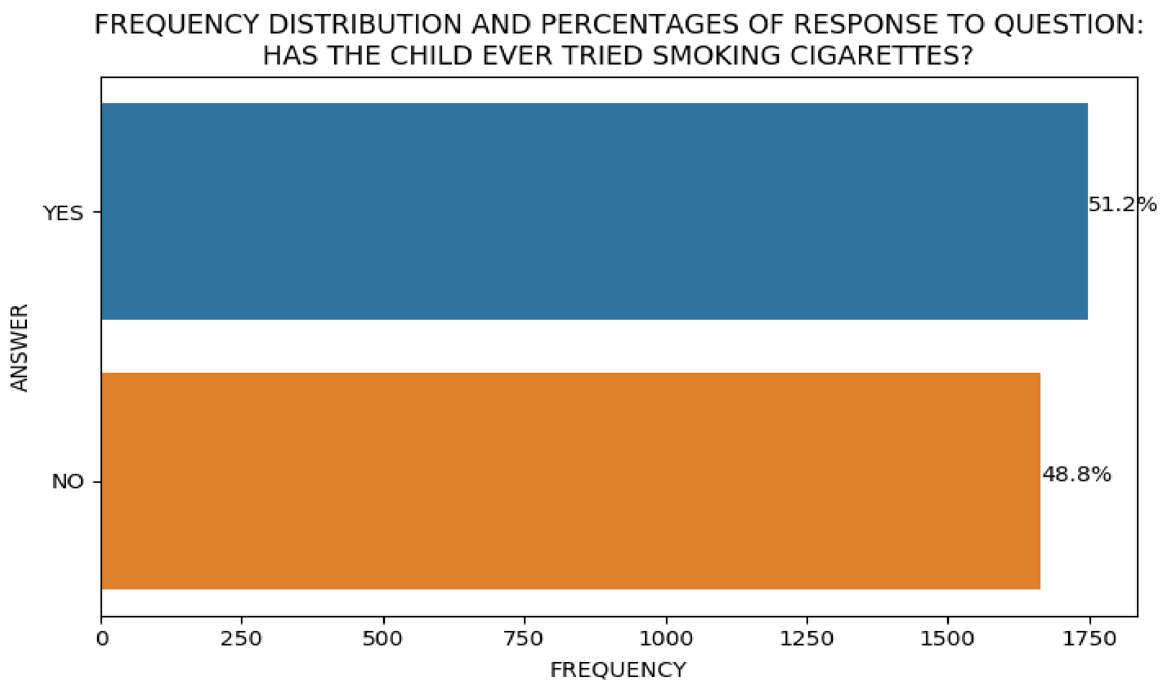
For displaying the average education level of parents, I binned it into 5 categories and gave them meaningful names. The binning operation was demonstrated earlier in this report. Originally the variable is numerical and after the binning, it was converted to categorical. The bar chart is horizontally oriented because this way the person viewing it can easily compare the values for each category. The results can be interpreted by saying that most parent body sets have attended or graduated from high school. The second most common type of education is university. Vocational school have scored closely to university graduates. Only a fraction of the parents has completed a form of education beyond a four-year college degree. The number of parents that have no education is such a small number that it's nearly insignificant.



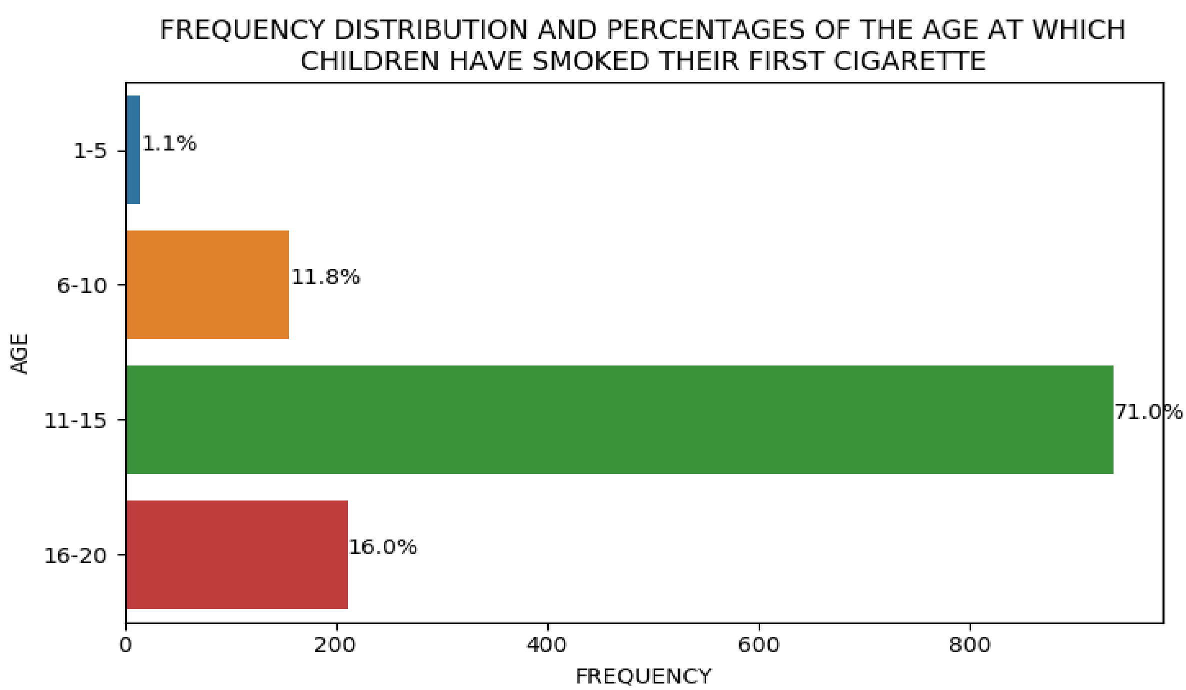
It was determined if children's parents are bossy or soft by collecting answers to questions from "***H1WP1***" through "***H1WP7***". These questions are of format "*Do your parents let you make your own decisions about...*". The possible answers are two – either yes, or no. If a set of parents let their children make their own decision about 4 or more questions (out of 7), they are soft. The chart displays both frequencies and percentages for each category. From the results one can tell that there are twice more soft parents than there are bossy ones.



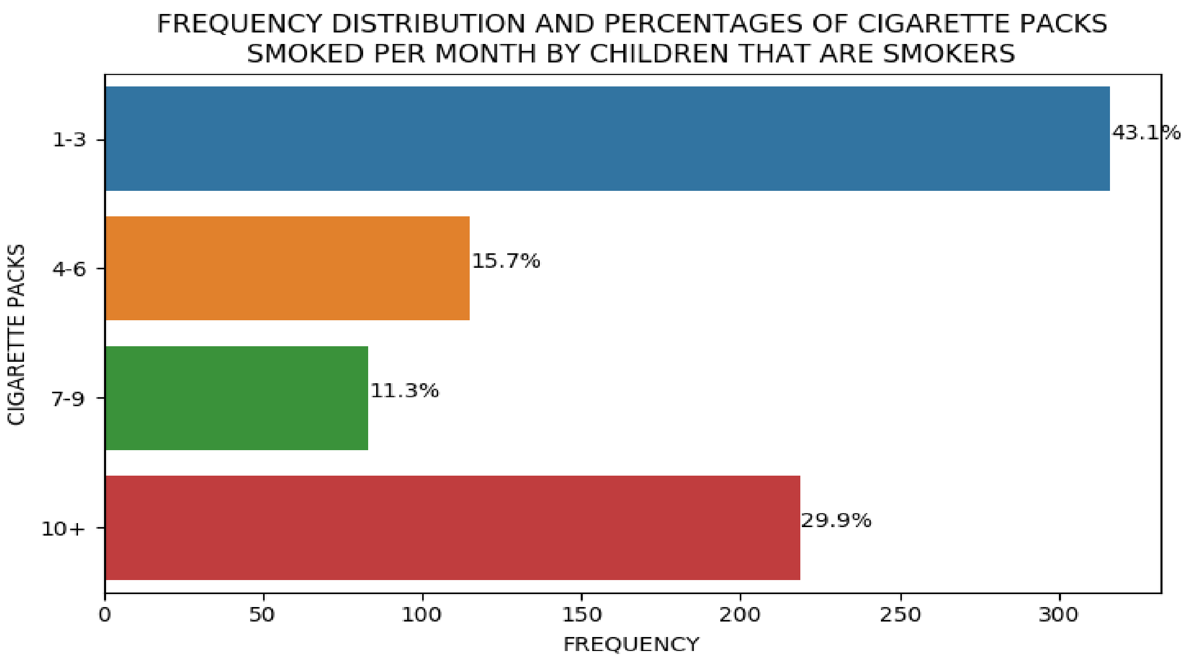
A bond score was calculated, determined by how close children are with their parents and how much they think their parents care about them. The score is of number format and it ranges from 1 (not at all) to 5 (very much). These scores are then binned into 3 categories (Low, Medium and High). The results can be interpreted by saying that almost all children have high bond with their parents and only a small portion of the children are in bad terms with their parent bodies.



This chart represents the frequency distribution and percentages for answers to question "Has the child ever tried to smoking cigarettes". It can be inferred from it that half the children have tried smoking, while the other half have not. However, slightly more participants have tried smoking cigarettes.



Officially, this dataset represents participants from 7th to 12th class in the USA. This means that they are aged between 12 to 18 years. From the obtained results we can tell that the people who answered the questionnaire have lied about their age. If we were to ignore that fact, the chart tells us that most of the children have tried smoking their first cigarette at age between 11-15. This age is amidst the teenage years and it's expected from adolescents to act chaotic.



This chart presents the average number of cigarette packs smoked by participants per month. From the given percentages one can tell that most of the children either smoke very small number of cigarette packs or a lot them. For this variable, the number of observations is naturally limited to those who smoke. My next step was to find out the mean, standard deviation and other descriptive statistics for the number cigarettes smoked per month:

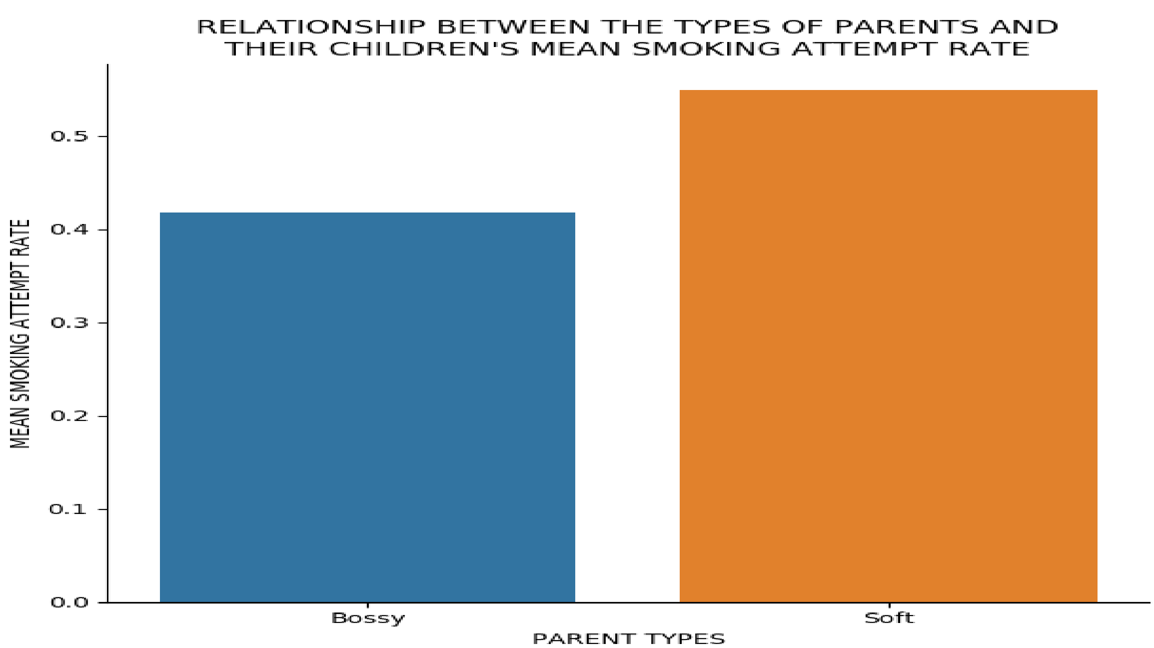
```
===============================================================================
Descriptive statistics about number of cigarettes smoked by smokers per month:

count     733.000
mean      194.555
std       248.677
min        30.420
25%        30.420
50%        91.260
75%       304.200
max      2707.380
Name: CIG_MONTHLY, dtype: float64
```

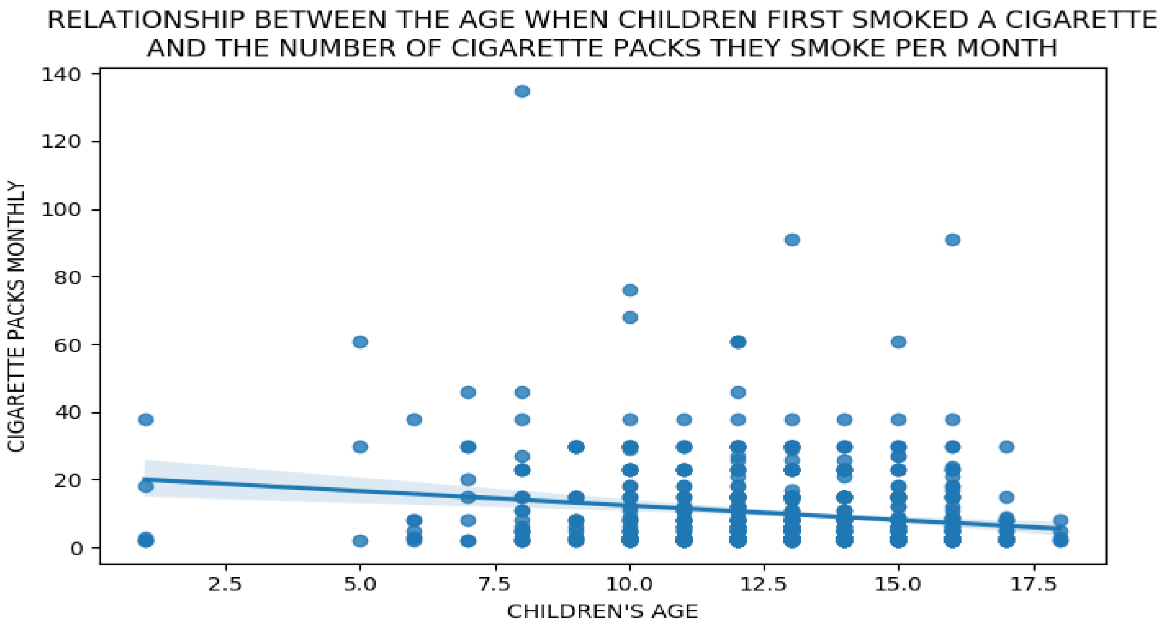
From the descriptive statistics one can tell that there are 733 children that are smokers. The average number of cigarettes they smoke per month is 195. The standard deviation value is close to the mean, which means that most of the data is concentrated in that portion. However, from the max value, one can tell that some students smoke a lot more than the average. This could be due to children not answering this question seriously. The minimum number of cigarettes smoked per month is 30, or 1 per day.

Bivariate

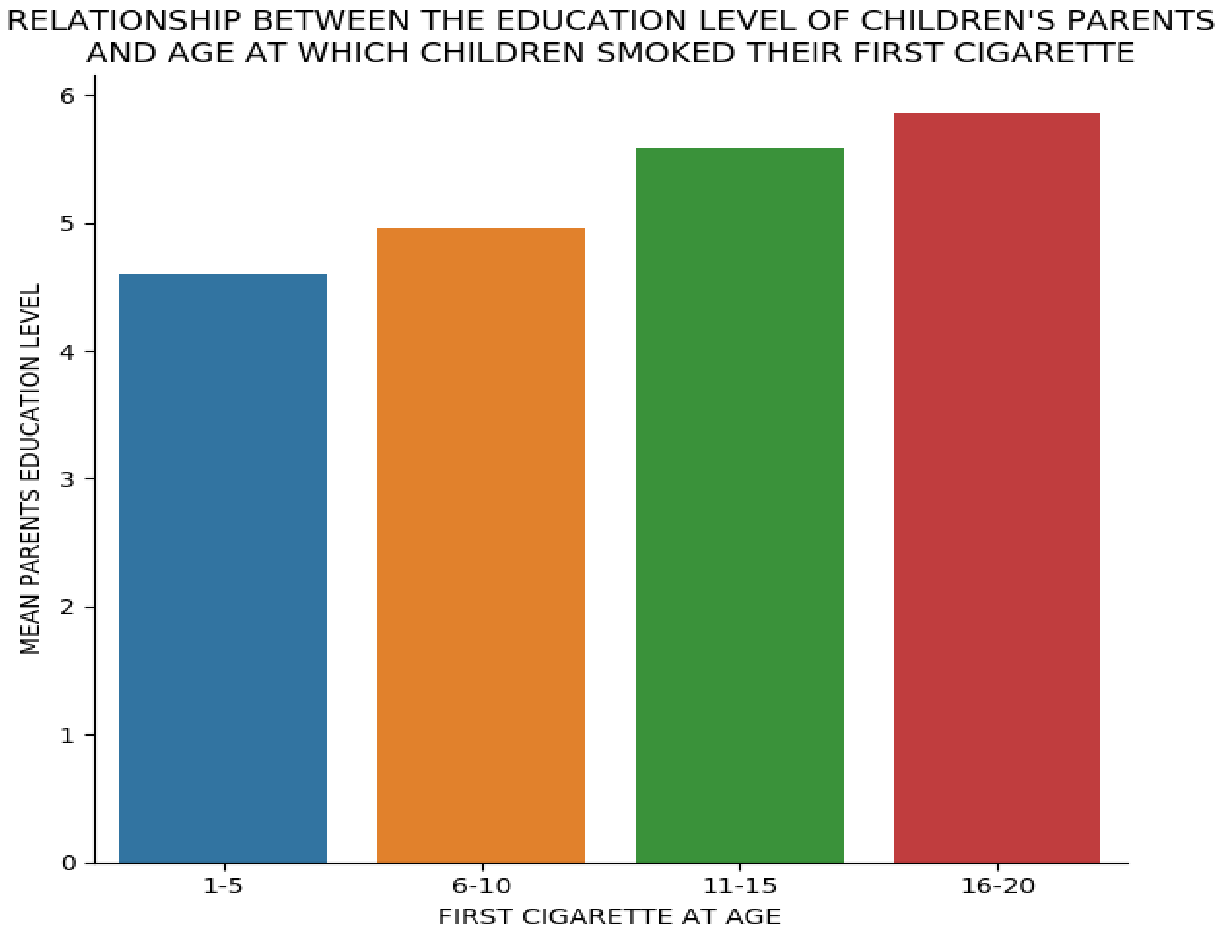
Without showing relationships between variables, findings do not possess great meaning. In this section I will show graphs that determine if an increase in one variable correlate with increase in another (or decrease) and whether high percentages in one category result in certain values of another variable.



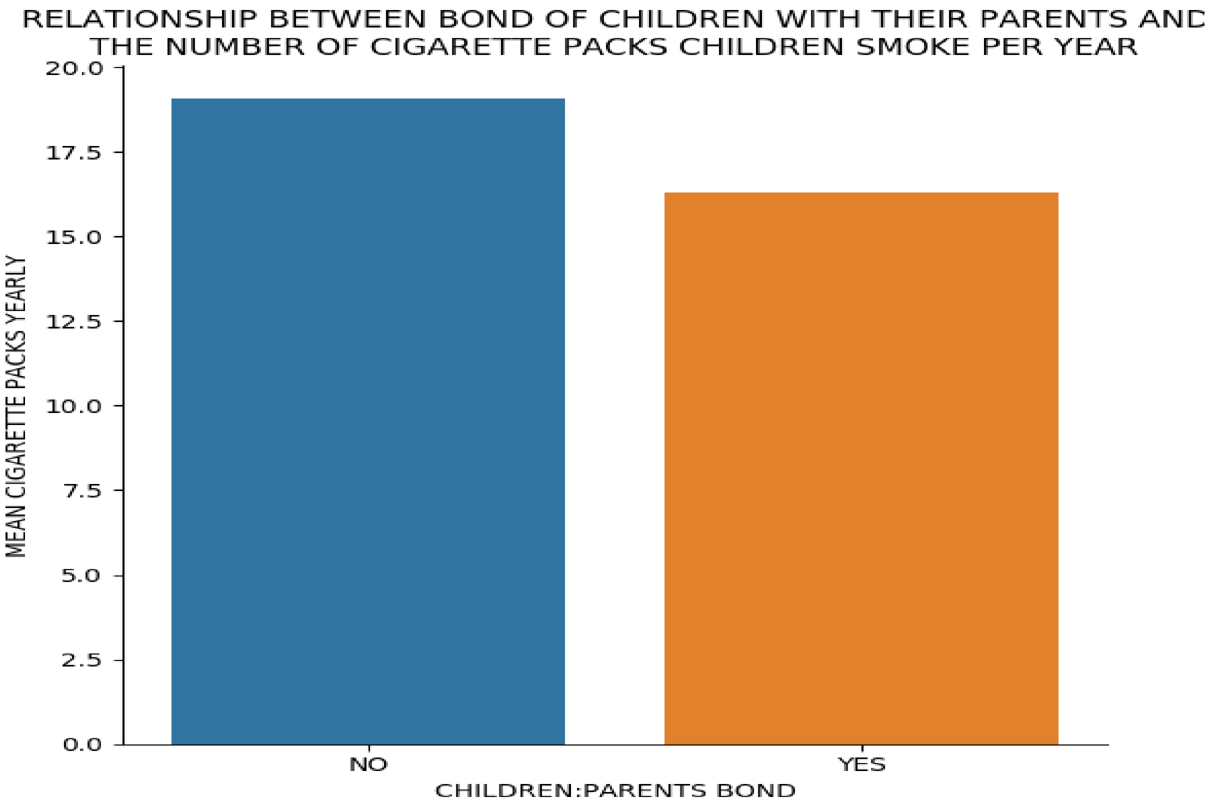
This graph has a response variable of mean smoking attempt rate (for children) and explanatory variable of parents' type. From the graph results, it can be said that children raised by soft parents have a higher chance of attempting to smoke cigarettes than children raised by bossy parents. The difference between the two groups is not dramatic, however it cannot be ignored.



This graph has an explanatory variable of children's age and a response variable of average number of cigarette packs smoked per month. It demonstrates the relationship between these two variables. The chart shows that most of children that smoke cigarettes use between 1 and 40 cigarette packs per month. There are outliers, which show that some participants smoke a lot more than the average number of cigarette packs per month. This could be because the participants of this study were not honest in their answers, which in results skewed the graph. Another important fact to note is that the number of packs rockets at the age of 10 and keeps its pace until the maximum age of 18. This chart does not prove any obvious relationship between the two variables.



This graph has an explanatory variable of age when first cigarette was smoked children and a response variable of average parents' education level. From the results one can say that children raised by parents with higher education level have bigger chance of attempting to smoke at a later age. This chart shows that there is a positive relationship between the age of first cigarette smoked by children and education level of their parents. The first two categories indicate a very early age, which indicates that the children that were answering questions for this study have lied about their age, as this study was conducted only on children aged between 12 and 18 years.



This graph has an explanatory variable of whether the children and their parents have a bond and a response variable of average number of cigarette packs smoked by children per year. The graph shows that children that do not have a bond with their parents are likely to smoke more cigarette packs per year and that children which have a bond with their parents – smoke less. The average number of cigarette packs smoked by children with no bond is 19, while for the other category is 16.5. The difference is not significant; however, it cannot be ignored.

Hypothesis Testing

- Were correlations identified and interpreted?
- What hypothesis testing was carried out and were the findings presented and interpreted?

Hypothesis testing is a statistical method used for making decisions about data. With hypothesis testing, one makes an assumption about population and proves it through statistical operations. In the end a conclusion is made. All hypothesis tests have two required parameters – null hypothesis and alternative hypothesis. In this section multiple hypothesis test will be performed. In the end conclusion is to be made. The types of tests used are Chi-Squared and ANOVA.

To make code cleaner and easier to use, I abstracted the hypothesis tests in their own functions.

```python
def chi2test(dataset, var_a, var_b, h0, h1, alpha=0.05):
    """
    Creates a contingency table for passed in variables and
    runs chi-squared test to determine whether to reject/keep null hypothesis.
    :param dataset: DataFrame
    :param var_a: str
    :param var_b: str
    :param h0: str
    :param h1: str
    :param alpha: float
    :return: None
    """
    print('\n=====================================================================

    # Create contingency table for passed in variables.
    crosstab = pd.crosstab(dataset[var_a], dataset[var_b])
    print(f'\n{crosstab}\n')

    # Chi-square test of independence of variables in a contingency table.
    print(f'\nRunning chi-squared test on variables \'{var_a}\' and \'{var_b}\':\n\n')
    stat, p, dof, expected = stats.chi2_contingency(crosstab)

    # Determine whether to reject or keep null hypothesis
    print(f'Significance: α = {alpha}\n'
          f'p-value: {p}\n'
          f'Degrees of freedom: {dof}\n'
          f'Expected: {expected}\n\n')

    if p <= alpha:
        print(f'Rejected H0: {h0}\n'
              f'The result is: {h1}\n')
    else:
        print(f'Failed to reject H0: {h0}\n')
```

```python
def anova(dataset, var_a, var_b, h0, h1, alpha=0.05):
    """
    Creates an OLS model model and determines whether
    to reject/keep null hypothesis based on the p-value.
    :param dataset: DataFrame
    :param var_a: str
    :param var_b: str
    :param h0: str
    :param h1: str
    :param alpha: float
    :return: None
    """
    # Using OLS function for calculating the F-statistic and associated p-value.
    model = smf.ols(formula=f'{var_a} ~ C({var_b})', data=dataset).fit()
    print(model.summary())

    p = model.pvalues[1]

    if p <= alpha:
        print(f'Rejected H0: {h0}\n'
              f'The result is: {h1}\n')
    else:
        print(f'Failed to reject H0: {h0}\n')

    print(
        "\n=====================================================================
        f'Means of {var_a} for all {var_b} categories:\n\n'
        f'{dataset.groupby(var_b).mean()}\n')
    print(
        "\n=====================================================================
        f'Standard deviations of {var_a} for all {var_b} categories:\n\n'
        f'{dataset.groupby(var_b).std()}\n')
```

Using these functions, I will run hypothesis tests on several variables in the dataset to determine which hypothesis to accept/reject. The functions will provide meaningful output based on the outcome p-value from carried out tests. In simple terms if the resulting p-value is less than or equal to alpha (which by default is 5%) the null hypothesis will be rejected. A typical function call for ***chi2test*** is like follows:

```python
"""
Hypothesis:
H0: The type of parents and whether their children have tried smoking cigarettes are independent (no association).
H1: The type of parents and whether their children have tried smoking cigarettes are dependent to each other.

Significance level: α = 0.05
"""
helpers.chi2test(dataset=dataset,
                 var_b='H1TO1',
                 var_a='PARENT_TYPES',
                 h0='The type of parents and whether their children have tried '
                    'smoking cigarettes are independent (no association).',
                 h1='The type of parents and whether their children have tried '
                    'smoking cigarettes are dependent to each other.')
```

```
==========================================================================================================


H1T01          0.0   1.0
PARENT_TYPES
Bossy          606   434
Soft          1058  1291


Running chi-squared test on variables 'PARENT_TYPES' and 'H1T01':


Significance: α = 0.05
p-value: 1.5786328372017779e-12
Degrees of freedom: 1
Expected: [[ 510.64030688  529.35969312]
 [1153.35969312 1195.64030688]]


Rejected H0: The type of parents and whether their children have tried smoking cigarettes are independent (no association).
The result is: The type of parents and whether their children have tried smoking cigarettes are dependent to each other.
```

H0: The type of parents and whether their children have tried smoking cigarettes are independent (no association).

H1: The type of parents and whether their children have tried smoking cigarettes are dependent to each other.

The result of this chi2 test is that null hypothesis was rejected because p-value was 0.00000000000158, which is less than the alpha value of 0.05. This means that the probability that we would get difference of this size when the null hypothesis is true is 0.000000000158%. The data provides significant evidence against the null hypothesis, so one can reject the null hypothesis and accept the alternative hypothesis.

```python
"""
Hypothesis:
H0: There is no relationship between the number of cigarette packs smoked per month
    between children aged under 13 and over 13.
H1: There is a relationship between the number of cigarette packs smoked per month
    between children aged under 13 and over 13.

Significance level: α = 0.05
"""

# Subset of only two variables.
subset = dataset.loc[:, ['CIG_PACKS_MONTHLY', 'H1TO2']]
subset.rename(columns={"H1TO2": "SMOKE_AGES"}, inplace=True)  # Rename column.

# Bin into 2 categories - younger and older than 13 years.
subset['SMOKE_AGES_CAT'] = pd.cut(subset['SMOKE_AGES'],
                                  [0, 12, 18],
                                  labels=['<13', '>=13'])

helpers.anova(dataset=subset,
              var_a='CIG_PACKS_MONTHLY',
              var_b='SMOKE_AGES_CAT',
              h0='There is no relationship between the number of cigarette packs smoked per month'
                 'between children aged under 13 and over 13.',
              h1='There is a relationship between the number of cigarette packs smoked per month'
                 'between children aged under 13 and over 13.')
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      CIG_PACKS_MONTHLY   R-squared:                       0.026
Model:                            OLS   Adj. R-squared:                  0.024
Method:                 Least Squares   F-statistic:                     19.15
Date:                Sun, 08 Mar 2020   Prob (F-statistic):           1.38e-05
Time:                        21:15:11   Log-Likelihood:                -2858.3
No. Observations:                 730   AIC:                             5721.
Df Residuals:                     728   BIC:                             5730.
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  12.3507      0.716     17.240      0.000      10.944      13.757
C(SMOKE_AGES_CAT)[T.>=13]  -4.0294      0.921     -4.377      0.000      -5.837      -2.222
==============================================================================
Omnibus:                      599.225   Durbin-Watson:                   2.063
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            16104.596
Skew:                           3.532   Prob(JB):                         0.00
Kurtosis:                      24.899   Cond. No.                         2.95
==============================================================================
```

```
Rejected H0: There is no relationship between the number of cigarette packs smoked per monthbetween children aged under 13
 and over 13.
The result is: There is a relationship between the number of cigarette packs smoked per monthbetween children aged under 13
 and over 13.



============================================================================================================
Means of CIG_PACKS_MONTHLY for all SMOKE_AGES_CAT categories:

                CIG_PACKS_MONTHLY  SMOKE_AGES
SMOKE_AGES_CAT
<13                        12.351      10.514
>=13                        8.321      14.504



============================================================================================================
Standard deviations of CIG_PACKS_MONTHLY for all SMOKE_AGES_CAT categories:

                CIG_PACKS_MONTHLY  SMOKE_AGES
SMOKE_AGES_CAT
<13                        14.801       2.086
>=13                       10.071       1.296
```

H0: There is no relationship between the number of cigarette packs smoked per month between children aged under 13 and over 13.

H1: There is a relationship between the number of cigarette packs smoked per month between children aged under 13 and over 13.

The result of this ANOVA test is that null hypothesis was rejected because p-value was 0.0000138, which is less than the alpha value of 0.05. This means that the probability that we would get difference of this size when the null hypothesis is true is 0.00138%. The data provides significant evidence against the null hypothesis, so one can reject the null hypothesis and accept the alternative hypothesis.

```python
"""
Hypothesis:
H0: The bond parents:children and number of cigarette packs children smoke are independent (no association).
H1: The bond parents:children and number of cigarette packs children smoke are dependent to each other.

Significance level: α = 0.05
"""
helpers.chi2test(dataset=dataset,
                 var_b='CIG_PACKS_YEARLY',
                 var_a='PARENTS_CHILD_BOND_OR_NOT',
                 h0='The bond parents:children and number of cigarette packs '
                    'children smoke are independent (no association).',
                 h1='The bond parents:children and number of cigarette packs '
                    'children smoke are dependent to each other.')
```

```
==========================================================================================================


CIG_PACKS_YEARLY           3.0    5.0    8.0    ...  127.0  152.0  226.0
PARENTS_CHILD_BOND_OR_NOT                       ...
NO                           2      1      0    ...      0      0      0
YES                        199    114     72    ...      1      2      1

[2 rows x 27 columns]


Running chi-squared test on variables 'PARENTS_CHILD_BOND_OR_NOT' and 'CIG_PACKS_YEARLY':


Significance: α = 0.05
p-value: 3.34189753561249e-06
Degrees of freedom: 26
Expected: [[3.29058663e+00 1.88267394e+00 1.17871760e+00 7.03956344e-01
  1.04774898e+00 3.11050477e-01 2.78308322e-01 1.80081855e-01
  3.27421555e-02 9.82264666e-01 3.27421555e-02 1.80081855e-01
  1.63710778e-02 4.91132333e-02 6.54843111e-01 1.63710778e-02
  4.91132333e-02 6.54843111e-02 3.27421555e-02 6.38472033e-01
  1.63710778e-01 4.91132333e-02 8.18553888e-02 1.63710778e-02
  1.63710778e-02 3.27421555e-02 1.63710778e-02]
 [1.97709413e+02 1.13117326e+02 7.08212824e+01 4.22960437e+01
  6.29522510e+01 1.86889495e+01 1.67216917e+01 1.08199181e+01
  1.96725784e+00 5.90177353e+01 1.96725784e+00 1.08199181e+01
  9.83628922e-01 2.95088677e+00 3.93451569e+01 9.83628922e-01
  2.95088677e+00 3.93451569e+00 1.96725784e+00 3.83615280e+01
  9.83628922e+00 2.95088677e+00 4.91814461e+00 9.83628922e-01
  9.83628922e-01 1.96725784e+00 9.83628922e-01]]


Rejected H0: The bond parents:children and number of cigarette packs children smoke are independent (no association).
The result is: The bond parents:children and number of cigarette packs children smoke are dependent to each other.
```

H0: The bond parents:children and number of cigarette packs children smoke are independent (no association).

H1: The bond parents:children and number of cigarette packs children smoke are dependent to each other.

The result of this chi2 test is that null hypothesis was rejected because p-value was 0.00000334, which is less than the alpha value of 0.05. This means that the probability that we would get difference of this size when the null hypothesis is true is 0.000334%. The data provides significant evidence against the null hypothesis, so one can reject the null hypothesis and accept the alternative hypothesis.

References

cpc.unc.edu. (n.d.). *Add Health*. [online] Available at: https://www.cpc.unc.edu/projects/addhealth [Accessed 8 Mar. 2020].

medium.com. (2019). *Introduction to Data Visualization in Python*. [online] Available at: https://towardsdatascience.com/introduction-to-data-visualization-in-python-89a54c97fbed [Accessed 8 Mar. 2020].