INTRODUCTION

SAS (Statistical Analysis System) is a powerful software suite for advanced analytics, multivariate analysis, business intelligence, data management, and predictive analytics. It allows the user to manipulate data, model statistics and visualize data using different procedures in the SAS software. In this tutorial, we will be looking at two procedures: PROC TABULATE and PROC SGPLOT.

1. OVERVIEW OF PROC TABULATE

PROC TABULATE is a crucial SAS procedure for statistical computation as it allows computation of various statistics and perform crosstabulations It is also useful for exploratory data analysis as it aids in discovering patterns and relationships in data, making it valuable for initial data exploration.

2. OVERVIEW OF PROC SGPLOT

PROC SGPLOT is a SAS procedure designed for statistical graphics and data visualization. Its purpose is to create high quality graphs and plots for visually exploring and communicating data patterns. It is important to effectively communicate analytical findings and improve interpretability of results through tailored high-quality graphics. These graphics aid in identifying relationships, patterns and outliers and allowing the user to have a deeper understanding of the data being explored.

3. DATASET OVERVIEW

The "Adult" dataset encompasses a diverse set of demographic and socioeconomic attributes, providing valuable insights for analysis. Key variables include:

- Age: Represents individuals' age in years.
- Education: Indicates the highest education level attained, ranging from Preschool to Doctorate.
- Sex: Represents the gender of individuals, categorized as Female or Male.
- HoursPerWeek: Indicates the average number of working hours per week.
- Income: Binary variable indicating whether an individual earns more than 50K or less than or equal to 50K.

EXPLANATION OF PROC TABULATE OPTIONS

1. KEY OPTIONS IN PROC TABULATE

1.1. CLASS Option:
- 1.1.1. Specifies categorical variables for defining rows and columns in the table.
- 1.1.2. Organizes data into categories, allowing crosstabulations and summary statistics based on these categories.

1.2. VAR Option:
- 1.2.1. Identifies numeric variables for analysis and summarization.
- 1.2.2. Specifies variables for analysis, providing summary statistics such as mean, sum, or other measures.

1.3. TABLE Option:
- 1.3.1. Defines the structure of the table, indicating how CLASS and VAR variables should be arranged.
- 1.3.2. Allows the creation of customized tables, specifying the arrangement of variables in rows and columns and the type of summary statistics.

1.4. SUMMARY Option:
- 1.4.1. Produces summary statistics for the variables listed in the VAR statement.
- 1.4.2. Calculates summary statistics like mean and sum for specified numeric variables.

1.5. TITLE Option:
- 1.5.1. Adds a title to the output.
- 1.5.2. Provides a clear description of the analysis being performed.

1.6. PRINTMISS Option:

1.6.1. Controls the display of missing values in the output.

1.6.2. When enabled, ensures that missing values are explicitly shown in the output, enhancing transparency in reporting.

2. Example SAS Code (PROC TABULATE)

2.1. CODE

```
proc tabulate data=your_lib.Adult; class education income; var hours_per_week; table education, income*hours_per_week*(mean) / printmiss; title 'Average Hours Worked per Week by Education and Income'; run;
```

2.2. EXPLANATION OF THE CODE:

- CLASS statement: Defines 'education' and 'income' as categorical variables.
- VAR statement: Identifies 'hour_per_week' as the numeric variable for analysis.
- TABLE statement: Determines the table structure
  - arranges 'education' in rows and 'income' in columns.
  - Calculates the mean of 'hours_per_week' for each combination.
- TITLE option: Adds a descriptive title to the output.
- PRINTMISS option: Ensures missing values are displayed as 'NA' in the output.

2.3. Interpretation of PROC TABULATE Output (Table 1.1)

- Individuals with higher education levels tend to work more hours on average especially in the >50K income category.
- In most cases, individuals with higher education levels tend to work more hours when their income is >50K compared to those with <=50K income which may mean that individuals with advanced degrees often work longer hours in higher paying occupations.
- The 'Preschool' category has a missing value for >50K income. This might be due to the dataset, where individuals with preschool education might not be common in high-income categories.

3. EXAMPLE SAS CODE FOR A COMPLEX TABLE USING PROC TABULATE

3.1. CODE

```
proc tabulate data=your_lib.Adult; class education income;var hours_per_week; table education, income*hours_per_week*(mean='Mean' sum='Sum' n='Count')* f=comma9.2 / printmiss; title 'Complex Table of Hours Worked per Week by Education and Income';run;
```

3.2. EXPLANATION OF THE CODE:

- CLASS statement: Organizes the table by 'education', nested within 'income'.
- VAR statement: Identifies 'hour_per_week' as the numeric variable for analysis.
- TABLE statement: Determines the table structure
  - ‘education’ will be the primary row of the table
  - defines the nested columns and statistics and shows mean, sum, and count for 'hours_per_week' for each combination of 'education' and 'income'
  - formats numeric values with commas and two decimal places.
- Title option: Adds a descriptive title to the output.

3.3. INTERPRETATION OF COMPLEX TABLE USING PROC TABULATE (TABLE 1.2)

- At lower education levels, the relationship between education and hours worked is less consistent. Table 1.2 shows those with 'Assocac', 'HSgrad', and 'Somecollege' qualifications have different patterns in terms of average hours worked.
- Across all education levels, individuals in the '>50K' income category generally work more hours than those in the '<=50K' category. This aligns with the common notion that higher income individuals often engage in longer working hours.
- The impact of income on hours worked varies across educational levels. While the average hours worked increases with income for most education levels, the extent of this increase differs.
- The 'Preschool' category shows a lower average hour worked, which is expected given the nature of this education level. The missing values at '>50k' suggests that there might be a lack of data or observations for individuals with preschool education earning '>50K'.

4. EXAMPLE SAS CODE FOR SUMMARY STATISTICS BY EDUCATION AND GENDER

4.1. CODE

```
proc tabulate data=your_lib.Adult;class education sex;var hours_per_week;

table education*(mean hours_per_week median hours_per_week), sex*(mean hours_per_week median hours_per_week);title 'Summary Statistics of Hours Worked per Week by Education and Gender';run;
```

4.2. EXPLANATION OF THE CODE:

- CLASS statement defines 'education' and 'sex' as categorical variables.
- VAR statement identifies the 'hour_per_week' as numeric variable to be analyzed.
- TABLE statement determines the structure of the table, education:
    - 'education' will be the rows of the table
    - Calculates the mean and median for each combination.
- TITLE ' Summary Statistics of Hours Worked per Week by Education and Gende e' Adds a title to the output, providing a clear description of the analysis being performed.
- PRINTMISS option ensures that missing values are displayed as 'NA' in the output.

4.3. EXPLANATION OF THE OUTPUT (TABLE 1.3)

- Table 1.3 shows that for each education level, individuals with an income >50K tend to have a higher mean and sum of hours worked per week compared to those with an income <=50K.
- As the education level increases, there is a general trend of an increase in the mean and sum of hours worked per week, irrespective of income category.
- The mean and sum of hours worked for individuals with a Doctorate and income >50K are notably higher than for other education levels.
- For the "Preschool" category, the mean and sum of hours worked are provided only for income <=50K, which suggests that individuals with education limited to preschool are less likely to fall into the >50K income category.

EXPLANATION OF PROC SGPLOT OPTIONS

1. KEY OPTIONS IN PROC SGPLOT

1.1. SCATTER Option:

1.1.1. Generates scatter plots.

1.1.2. Visualizes the relationship between two continuous variables, coloring points based on a categorical variable.

1.2. VBAR Option:

1.2.1. : Generates a vertical bar chart.

1.2.2. : Visualizes the relationship between a categorical and numerical variable.

1.3. TITLE Option:

1.3.1. : Adds a title to the plot.

1.3.2. : Provides a clear and concise title, improving interpretability.

1.4. XAXIS and YAXIS Options:

1.4.1. : Adds labels to the x-axis and y-axis.

1.4.2. : Provides context to the variables being plotted.

1.5. LEGEND Option:

1.5.1. : Adds a legend to the plot.

1.5.2. : Identifies different groups in the plot.

1.6. HBOX Option:

1.6.1. : Generates box plots.

1.6.2. : Visualizes the distribution of a numerical variable, showing median, quartiles, and outliers.

2. SCATTER PLOT OF AGE AND HOURS PER WEEK BY INCOME

2.1. CODE

```
proc sgplot data=your_lib.Adult; scatter x=age y=hours_per_week / group=income; title 'Scatter Plot of Age and Hours per Week by Income'; xaxis label='Age'; yaxis label='Hours per Week'; legend 'Income' / title='Income Categories';run;
```

2.2. EXPLANATION OF THE CODE:

- SCATTER visualizes the relationship between two continuous variables (x=age and y=hours_per_week in this case) and colors the points based on a categorical variable (group=income).
- TITLE ' Scatter Plot of Hours Worked by Education and Income ' adds a title to the output, providing a clear description of the analysis being performed
- XAXIS and YAXIS add labels to the axes, providing context to the variables being plotted.
- LEGEND identifies income categories in the plot
- Education levels are shown on the xaxis, and hours worked per week are on the yaxis.

2.3. INTERPRETATION OF THE CODE (Table 2.1)

- Unlike certain professions where work hours might be directly correlated with age (e.g., longer hours for younger individuals), this dataset does not exhibit a straightforward relationship.
- The absence of a clear pattern in the scatterplot suggests that factors other than age significantly contribute to the variation in hours worked.
- While there's variation in work hours, the distinction between income categories isn't as pronounced.
- Data point clusters may indicate potential subgroups within the dataset that share similar age and work hour characteristics, but more analysis is needed to understand the factors contributing to these clusters.

3. VERTICAL BAR CHART

3.1. CODE

```
proc sgplot data=your_lib.Adult; vbar education / group=income response=hours_per_week stat=mean;
title 'Average Hours Worked per Week by Education Level and Income';
xaxis label='Education Level'; yaxis label='Average Hours per Week'; run;
```

3.2. EXPLANATION OF THE CODE:

- VBAR statement creates vertical bar chart with education levels on the xaxis, and the height of the bars representing the average hours worked per week. The bars are grouped by the income variable.
- The 'GROUP=INCOME' option in var is used to distinguish bars for the two different income groups: <=50K and >50K.
- The 'STAT=MEAN' option calculates and displays the mean (average) of the hours_per_week variable for each category of education and income.
- 'XAXIS' and 'YAXIS' statements set labels for the xaxis and yaxis, respectively.
- The 'TITLE' statement provides a title for the chart.
- The 'GROUP INCOME' / 'DISCRETELEGEND' statement adds a legend to the chart, providing a visual reference for the income groups.

3.3. EXPLANATION OF OUTPUT (TABLE 2.2)

- The vertical bar chart in figure 2.2 shows that individuals with higher education levels, such as 'Doctorate', 'Profschool', and 'Masters', tend to work more hours per week on average, especially in the '>50K' income category.
- The vertical bar chart shows that the "Profschool" and "Masters" category have taller bars for both income groups which suggests that individuals with professional schooling tend to work longer hours on average, regardless of their income level.
- There 'Preschool' bar shows a lower average hour worked, which is expected given the nature of this educational level but the missing values at '>50k' suggests that there might be a lack of data for individuals with preschool education earning '>50K'.

4. BOX PLOT OF HOURS WORKED PER WEEK BY EDUCATION LEVEL AND GENDER (ABOVE GRADE 12)

4.1. EXAMPLE SAS CODE FOR BOX PLOT

```
proc sgplot data=your_lib.Adult; title 'Box Plot of Hours Worked per Week by Education Level and
Gender (Above Grade 12)'; xaxis label='Education Level'; yaxis label='Hours per Week';hbox
hours_per_week / category=education group=sex;where education_num > 12; run;
```

4.2. EXPLANATION OF THE CODE:

- TITLE 'Box Plot of Hours Worked per Week by Gender' Sets the title of the plot.
- 'XAXIS LABEL' Configures the Xaxis label as "Gender."
- 'YAXIS LABEL': Configures the Yaxis label as "Hours per Week."
- 'HBOX hours_per_week / category = education group = sex'
    - hbox: Stands for horizontal box plot.
    - hours_per_week: Specifies the variable to be plotted on the Yaxis (hours worked per week).
    - category = education: Categorizes the data by education level on the Xaxis.
    - group = sex: Groups the data by gender. Separate box plots will be created for males and females within each education level.

4.3. EXPLANATION OF OUTPUT (TABLE 2.3)

- For each education level and sex, the box plot shows the median (line inside the box), interquartile range (box), and any outliers (points beyond the whiskers).
- The median hours worked per week generally increases with higher education levels for males.
- Males tend to have higher median hours worked compared to females in all education levels except bachelors.
- As seen in figure 2.2, there are lots of outliers in the data but for all groups except females at doctorate and prof school level have more outliers towards the left which means that hours worked per week are unusually low for everyone except the females at prof school and doctorate level.
- The females at doctorate level have a very wide interquartile range which could mean that the hours worked per week vary more for this group compared to other groups which could be due to a variety of reasons, such as differences in work schedules, hours of work, or types of jobs
- The interquartile range is generally narrower for females compared to males except at a doctorate level suggests that, within all education level except doctorates, females tend to cluster around the median more closely than males.

SUMMARY

1. **PROC TABULATE ANALYSIS**
   - 1.1. 1. AVERAGE HOURS WORKED BY EDUCATION AND INCOME:
     - 1.1.1. Higher education levels correlate with increased average hours worked.
     - 1.1.2. Individuals with advanced degrees tend to work longer hours, especially in the '>50K' income category.
   - 1.2. 2. COMPLEX TABLE OF HOURS WORKED BY EDUCATION AND INCOME:
     - 1.2.1. Confirms the positive correlation between higher education levels and longer working hours.
     - 1.2.2. Different patterns in average hours worked are observed across various education levels.
   - 1.3. 3. SUMMARY STATISTICS BY EDUCATION AND GENDER:
     - 1.3.1. For each education level, individuals with an income >50K exhibit higher mean and sum of hours worked.
     - 1.3.2. Mean and sum of hours worked generally increase with education level, particularly for individuals with a Doctorate and income >50K.

2. **PROC SGPLOT ANALYSIS:**
   - 2.1. SCATTER PLOT OF AGE AND HOURS PER WEEK BY INCOME:
     - 2.1.1. No clear pattern between age and hours worked is identified.
     - 2.1.2. Age alone does not explain the variation in working hours.
   - 2.2. VERTICAL BAR CHART AVERAGE HOURS WORKED PER WEEK BY EDUCATION AND INCOME:
     - 2.2.1. Higher education levels, especially 'Doctorate' and 'Profschool,' are associated with longer average hours worked.
     - 2.2.2. Professional schooling shows a positive correlation with increased working hours, particularly in the '>50K' income category.
   - 2.3. BOX PLOT OF HOURS WORKED PER WEEK BY EDUCATION LEVEL AND GENDER:
     - 2.3.1. Males generally have higher median hours worked compared to females, except at the bachelor level.
     - 2.3.2. Outliers are observed, particularly for females at doctorate and prof school levels.

3. **KEY FINDINGS ABOUT ADULT DATASET:**
   - 3.1. EDUCATION AND HOURS WORKED:

3.1.1. Higher education levels are linked to longer working hours, especially for advanced degrees.

3.1.2. 'Doctorate' and 'Profschool' show the strongest positive correlation with increased hours worked.

3.2. INCOME AND HOURS WORKED:

3.2.1. Individuals with income >50K generally work more hours, indicating a potential association between higher income and longer working hours.

3.3. GENDER DIFFERENCES:

3.3.1. Gender disparities exist, with males tending to work more hours than females.

3.3.2. Outliers for females at advanced education levels (above grade 12) warrant further investigation.

3.4. PRESCHOOL EDUCATION:

3.4.1. Preschool education is associated with lower average hours worked.

3.4.2. Missing values for '>50K' income suggest a potential lack of representation in higher income categories.

3.5. AGE AND HOURS WORKED:

3.5.1. Age alone does not explain variations in working hours.

3.5.2. Other factors beyond age significantly contribute to the observed patterns.

4. CONCLUSION

In summary, the analysis underscores the importance of education, income, and gender in understanding variations in working hours. Advanced education levels and higher income categories are generally associated with longer working hours, with gender differences and outliers contributing to the complexity of the relationships. Further exploration and multivariate analysis are recommended to get a deeper understanding of the factors affecting hours worked per week as type of occupation, job (part time, freelance...). These factors could explain the disparities in the data.

APPENDIX

1. PROC TABULATE TABLES

Average Hours Worked per Week by Education and Income

	income	
	<=50K	>50K
	hours_per_week	hours_per_week
	Mean	Mean
education		
10th	36.57	43.77
11th	33.32	45.13
12th	35.04	44.82
1st-4th	37.86	48.83
5th-6th	38.54	46.00
7th-8th	38.83	47.50
9th	37.67	44.85
Assoc-ac	39.26	44.26
Assoc-vo	40.82	43.85
Bachelor	40.59	45.48
Doctorat	45.43	47.51
HS-grad	39.73	45.04
Masters	41.22	45.92
Preschoo	36.65	.
Prof-sch	42.82	49.09
Some-col	37.45	44.82

Table 1.1: Average Hours Worked Per Week by Income and Education

Complex Table of Hours Worked per Week by Education and Income

	income					
	<=50K			>50K		
	hours_per_week			hours_per_week		
	Mean	Sum	Count	Mean	Sum	Count
education						
10th	36.57	31,856.00	871.00	43.77	2,714.00	62.00
11th	33.32	37,155.00	1,115.00	45.13	2,708.00	60.00
12th	35.04	14,014.00	400.00	44.82	1,479.00	33.00
1st-4th	37.86	6,134.00	162.00	48.83	293.00	6.00
5th-6th	38.54	12,217.00	317.00	46.00	736.00	16.00
7th-8th	38.83	23,531.00	606.00	47.50	1,900.00	40.00
9th	37.67	18,344.00	487.00	44.85	1,211.00	27.00
Assoc-ac	39.26	31,490.00	802.00	44.26	11,728.00	265.00
Assoc-vo	40.82	41,675.00	1,021.00	43.85	15,831.00	361.00
Bachelor	40.59	127197.00	3,134.00	45.48	101001.00	2,221.00
Doctorat	45.43	4,861.00	107.00	47.51	14,539.00	306.00
HS-grad	39.73	350635.00	8,826.00	45.04	75,447.00	1,675.00
Masters	41.22	31,495.00	764.00	45.92	44,035.00	959.00
Preschoo	36.65	1,869.00	51.00	.	.	0.00
Prof-sch	42.82	6,551.00	153.00	49.09	20,766.00	423.00
Some-col	37.45	221106.00	5,904.00	44.82	62,166.00	1,387.00

Table 1.2: Complex table of Hours Worked Per Week by Income and Education

Summary Statistics of Hours Worked per Week by Education and Gender

	sex			
	Female		Male	
	hours_per_week		hours_per_week	
	Mean	Median	Mean	Median
education				
10th	32.11	37.00	39.34	40.00
11th	29.82	30.50	36.31	40.00
12th	31.79	35.00	37.77	40.00
1st-4th	31.98	37.50	40.62	40.00
5th-6th	36.05	40.00	39.86	40.00
7th-8th	36.20	40.00	40.41	40.00
9th	33.92	40.00	39.65	40.00
Assoc-ac	37.36	40.00	42.55	40.00
Assoc-vo	37.83	40.00	43.75	40.00
Bachelor	39.33	40.00	44.04	40.00
Doctorat	47.30	40.00	46.89	45.00
HS-grad	36.58	40.00	42.48	40.00
Masters	41.11	40.00	45.07	42.00
Preschoo	31.88	35.00	38.83	40.00
Prof-sch	44.79	40.00	47.93	50.00
Some-col	34.57	40.00	41.53	40.00

Table 1.3 Summary Statistics of Hours Worked per Week by Education and Gender

2. PROC SGPLOT

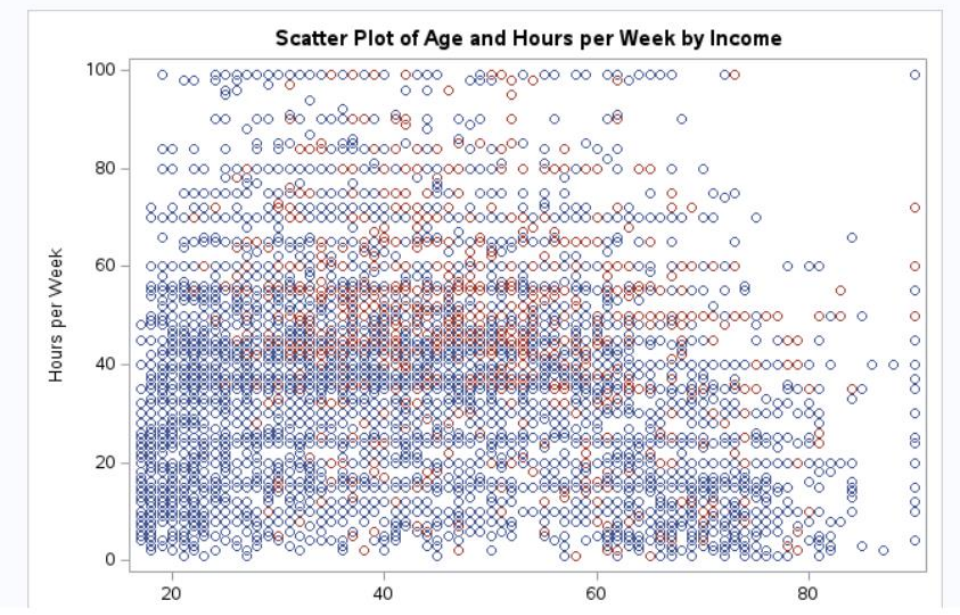

Table 2.1: Scatterplot of Age and Hours Worked Per Week by Income

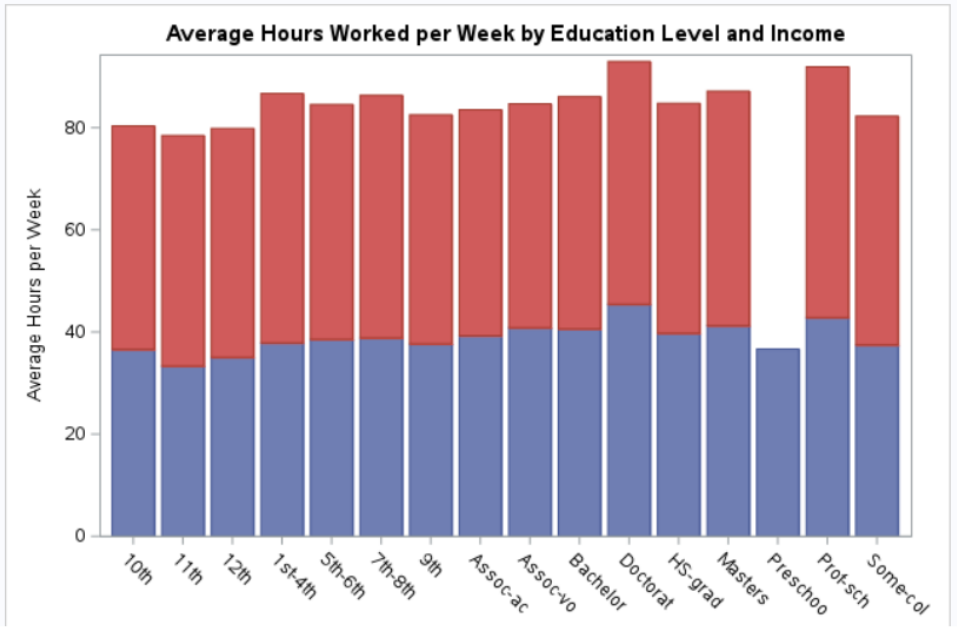

Table 2.2: Vertical Bar Chart of Average Hours Worked Per Week by Education Level and Income

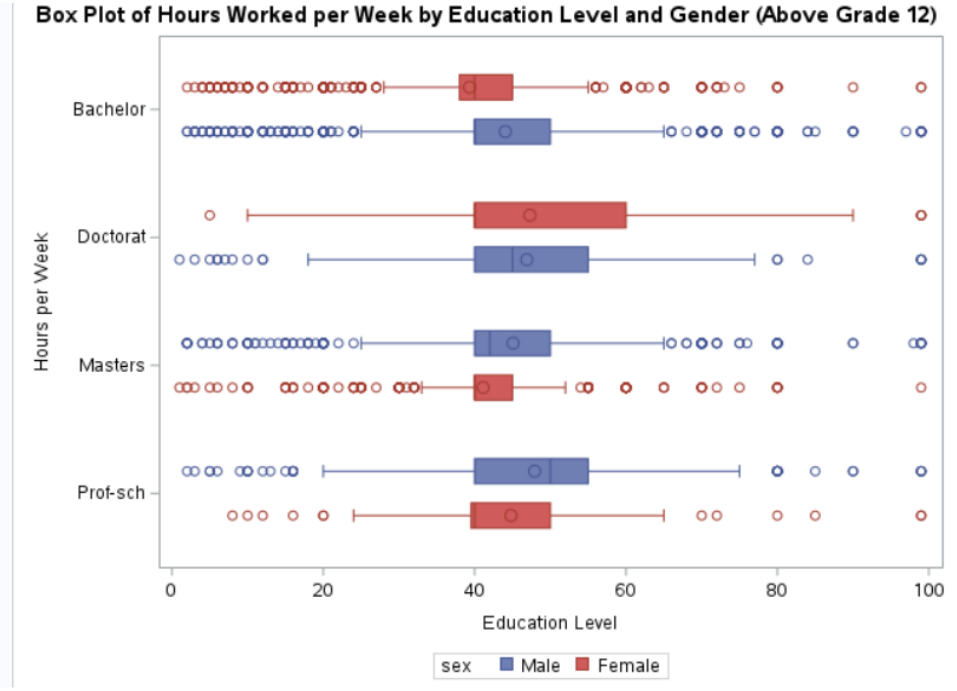

Table 2.3: Box Plot of Hours Worked per Week by Education Level and Gender (Above Grade 12)