MANCHESTER METROPOLITAN UNIVERSITY

MSC DATA SCIENCE

6G7Z1020

COMPUTATIONAL STATISTICS AND VISUALISATIONS

ASSIGNMENT 1: -
MELBOURNE CUP & SEXUAL ACTIVITY

DANIEL RAY
18053479

28/10/18

INTRODUCTION

In the first assignment of the Computational Statistics and Visualisation module, I was tasked with visualising two dataset. The first dataset being the Winning times of the famous Horse Races in Melbourne, Australia spanning from 1861 to 2013. The second dataset consists of a subset of the number of partners for 1682 males and 1850 females from a 1989-1991 US survey.

To visualise these datasets, I used a data analysis software called SAS using its own programming language, SAS Language, to input data files and output results of statistical analysis.

MELBOURNE CUP

The first question of the assignment asks to plot Time as a function of the Year regarding the Winning times of the famous Horse races in Melbourne. We are told that the data spans from 1861 to 2013. We can use SAS to produce a contents table so that we can examine the variables in the dataset in more depth:

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
2	Time	Num	8	BEST12.	BEST32.
1	Year	Num	8	BEST12.	BEST32.

Table 1 - Contents Table for Melbourne Cup Dataset

Looking at the contents table we can see there are two variable, Time and Year, which are both Number type variables of length 8. We can use SAS to further describe the dataset in more depth, using statistical operations:

Var	N	Mean	Std Dev	Minimum	Maximum
Time	153	206.8581699	7.0989556	196.3000000	232.0000000
Year	153	1937.00	44.3113981	1861.00	2013.00

Table 2 - Means Table for Melbourne Cup Dataset

In this table, we can see that the dataset has 153 entries, with the range of times spanning from 196.3 seconds to 232 seconds. We can see that the average time is 205.85 seconds, with a standard deviation of 7.1. As stated before but confirmed by the table, the dataset spans from 1861 to 2013.

Although being rather informative, both summary tables lack the ability to depict any trends within the dataset. Using SAS, we can plot the two variables on a graph to gain a more visual and informative representation of the dataset. As asked in the question, I will plot time as a function of the year:

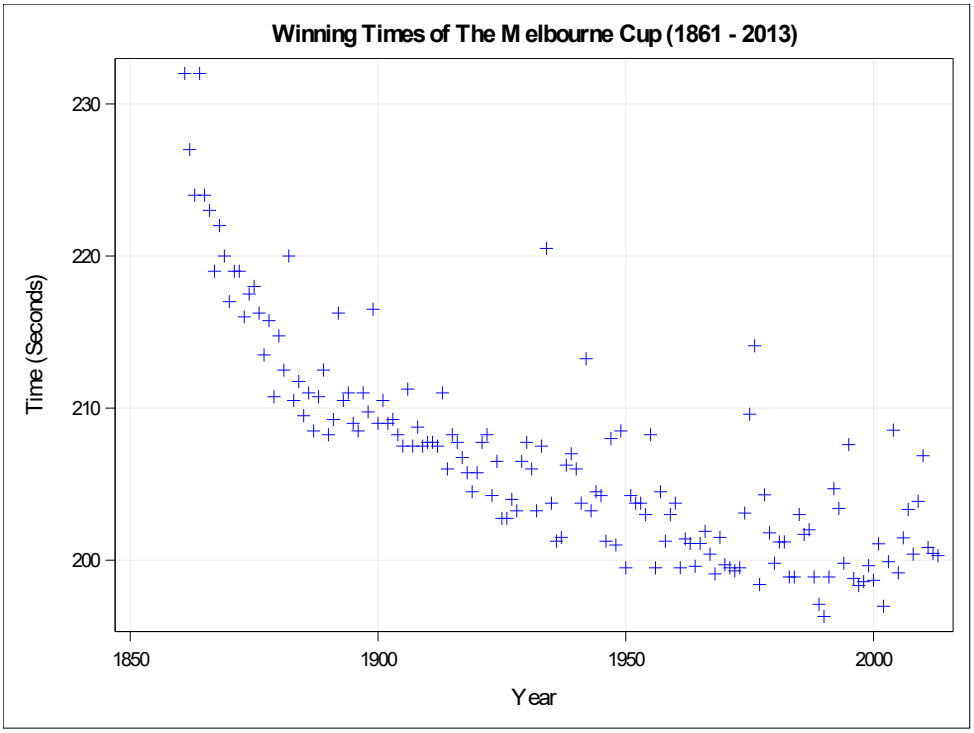

Figure 1 - Scatter Graph showing Time as a function of Year

As you can see, representing the dataset in a more visual method allows us to gain more information from the data. Upon seeing the graph, we can conclude that as the years increase the time of the winning races decrease. To gain a clearer picture of the relationship of the two variables in the dataset we can employ statistical models such as the linear regression model. This is a method for analysing the relationship between two quantitative variables, X and Y.

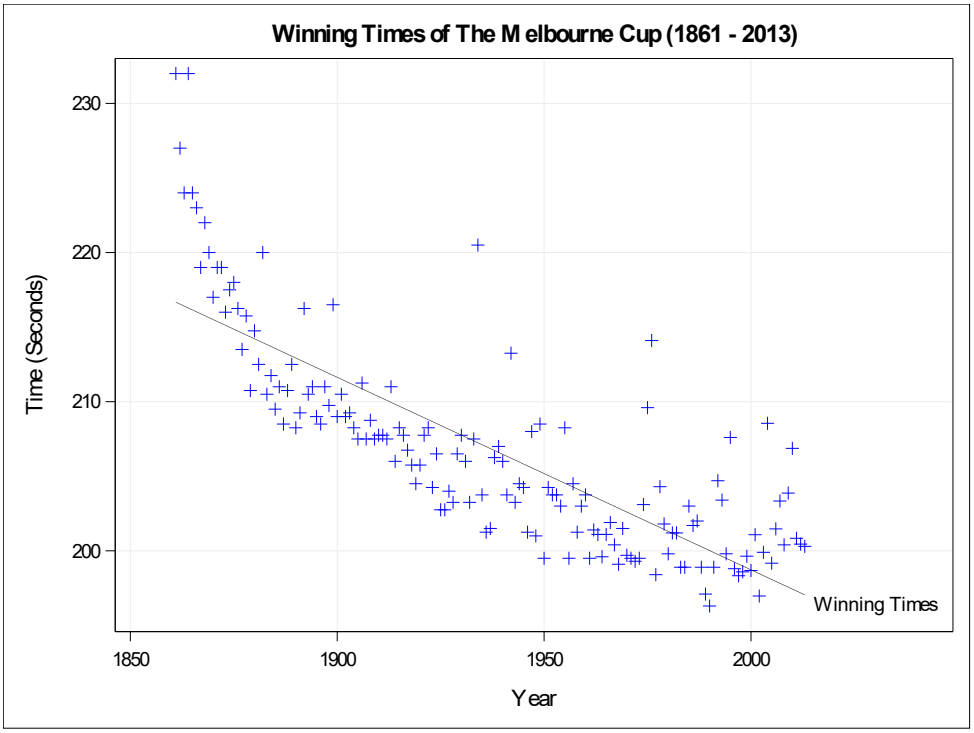

Figure 2 - Scatter Graph showing Linear Regression of Time as a function of Year

Now with the ability to see a clearer representation of the relationship between the two variables, we can suggest with a greater level of confidence that as the years increase the time decreases, for example, in the year 1900, the winning time would have roughly been 212 seconds and in the year 2000 the winning time would have roughly been 198 seconds.

Finally, to gain a higher defined representation of the dataset, we can employ another statistical model called the Loess regression. This model is a nonparametric technique that uses local weighted regression to fit a smooth curve through points in a scatter plot.

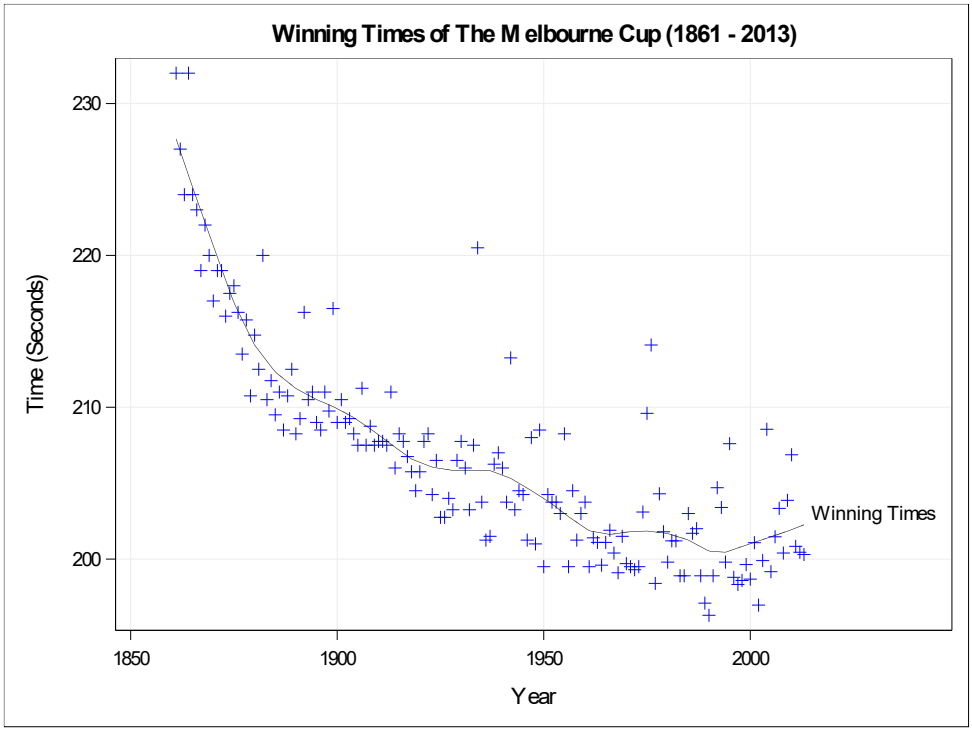

Figure 3 - Scatter Graph showing Loess Regression of Time as a function of Year

As expected, the relationship still stands that as the year increases the time decreases, however, looking at the loess regression we can see that from 1990 onwards, there has been a slight increase in winning time going from 201 to 203, which was not as apparent in the previous two graphs. Using the Loess Regression line, we can say that in 1900 the winning time would have roughly been 210 seconds and in 2000 the winning time would have roughly been 201.

CODE

```
/* Direct SAS to file location */
FILENAME REFFILE '/folders/myfolders/1/data/melb.csv';
/* Import dataset into SAS */
PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=melb;
GETNAMES=YES;
RUN;
ODS SELECT VARIABLES;
/*Produce the contents table of the dataset*/
PROC CONTENTS DATA=melb; RUN;
ODS SELECT DEFAULT;
/*Produce the stat desc of the dataset*/
PROC MEANS DATA=melb;
VAR TIME YEAR;
RUN;
/*Plot the scatter graph*/
PROC SGPLOT DATA = melb;
SCATTER X=Year Y=Time / markerattrs=(color=blue symbol=plus
size=10);
XAXIS GRID;
YAXIS GRID;
TITLE 'Winning Times of the Melbourne Cup (1861-2013)';
LABEL Time = 'Time (Seconds)';
RUN;
/*Plot the scatter graph w/ linear regression*/
PROC SGPLOT DATA = melb;
REG X=Year Y=Time / markerattrs=(color=blue symbol=plus size=10)
lineattrs=(color=black)
                              curvelabel='Winning Times' nolegfit;
XAXIS GRID;
YAXIS GRID;
TITLE 'Winning Times of the Melbourne Cup (1861-2013)';
LABEL Time = 'Time (Seconds)';
RUN;
/*Plot the scatter graph w/ loess regression(non-linear)*/
PROC SGPLOT DATA = melb;
LOESS X=Year Y=Time  / markerattrs=(color=blue symbol=plus size=10)
lineattrs=(color=black)
                              curvelabel='Winning Times' nolegfit;
XAXIS GRID;
YAXIS GRID;
TITLE 'Winning Times of the Melbourne Cup (1861-2013)';
LABEL Time = 'Time (Seconds)';
RUN;
```

SEXUAL ACTIVITY

In the second question of the assignment we are given a dataset containing two variables, Partners and Sex; where Partners are the total number of sexual partners that person has had along with their Sex, which is either Male (1682 entries) or Female (1850 entries). We have been given the task of plotting the data into a bar chart with the aim of identifying any similarities and differences. We can use SAS to examine the contents of the dataset:

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
1	Partners	Char	5	$5.	$5.
2	Sex	Char	3	$3.	$3.

Table 3 - The Contents Table of the SA Dataset

As expected, there are two variables – Partners and Sex, which are both Character type variables. In order to gain a richer understanding of the dataset, we can apply statistical operations to the variables. However, as the Partners variable is non-numeric SAS is unable to compute these values. I therefore created a new dataset containing a new variable which has a variable that contains the numerical value of the partners variable:

Alphabetic List of Variables and Attributes					
#	Variable	Type	Len	Format	Informat
3	INT_PARTNERS	Num	8		
1	Partners	Char	5	$5.	$5.
2	Sex	Char	3	$3.	$3.

Table 4 - The Contents Table of the New Dataset

Now that we have the numerical value of the number of partners variable, we can gain a statistical description of the variable:

Analysis Variable : INT_PARTNERS				
N	Mean	Std Dev	Minimum	Maximum
3532	7.3207814	17.9457595	0	253.0000000

Table 5 - Means Table for Partners Variable

We can now see that the average number of partners for both male and females is 7.3, with a standard deviation of 17.9. A minimum value of 0 and maximum value of 253. However, this doesn't allow us to compare the two sexes. We can use SAS to create a bar chart that shows the average number of partners for each sex:

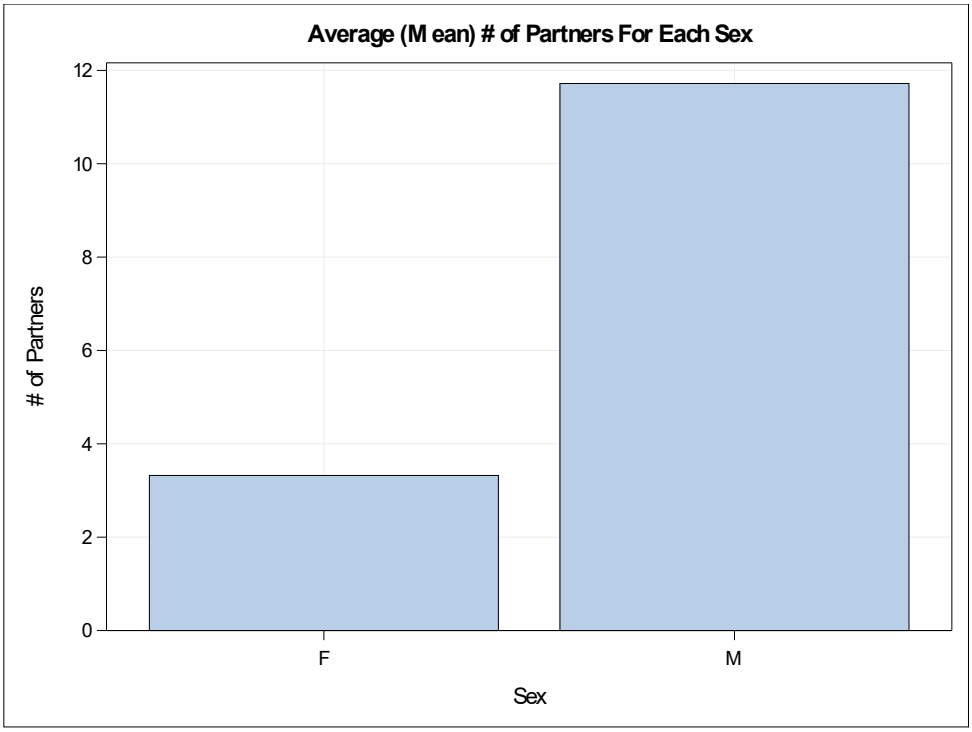

Figure 4 - Bar Chart of the Mean # of Partners for Each Sex

Here we can see that on average Females have 3.6 sexual partners whilst Males have on average 11.8 partners. Although informative, the graph above doesn't depict the whole of the dataset. We can use SAS to produce a bar chart that allows us to see how the data set is distributed for each sex. First we will examine the Males:

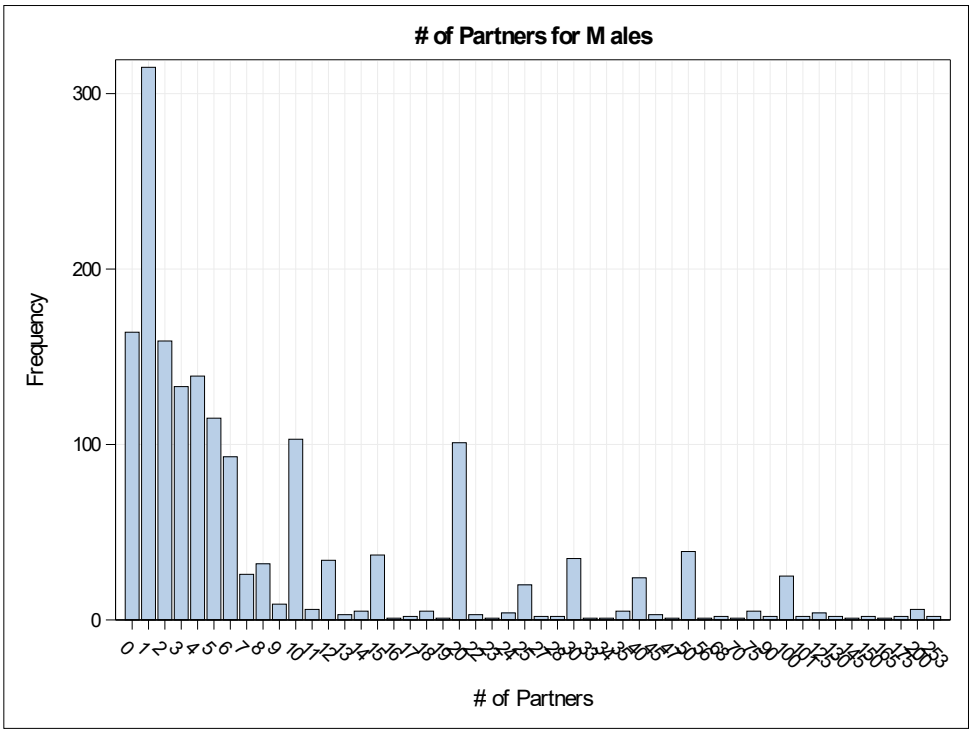

Figure 5 - Bar Chart of Frequencies for the # of Sexual Partners of Males

As we can see the bar chart depicts the frequencies for the number of partners for males. The highest frequency of partners for males is 1, with more than 300 males. We can also see that the majority of males have 10 or less sexual partners but there is a large spread, with the highest number of partners being 253. Using SAS we can now do the same for Females:

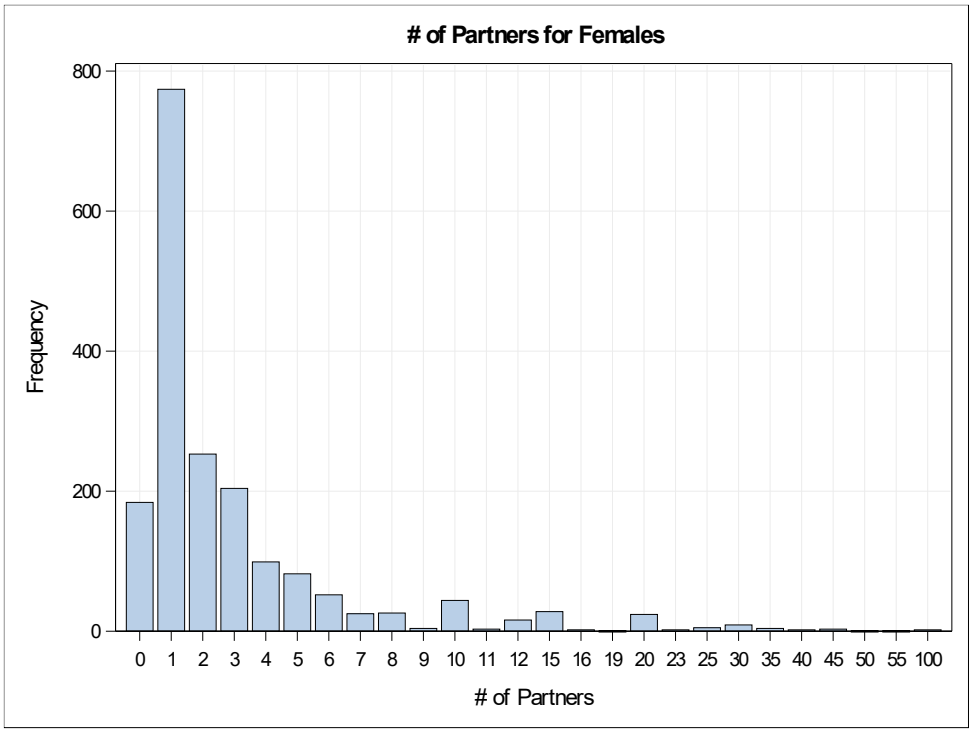

Figure 6 - Bar Chart of Frequencies for the # of Sexual Partners of Females

Looking at the frequencies for the number of sexual partners of females we can see that similar to the males, 1 sexual partner is the most frequent number of sexual partners with nearly 800 females having only one sexual partner. Like the males, the majority of females have 10 or less partners. Unlike the males however, the spread of sexual partners is smaller for females with the highest number of partners being 100.

Finally, we can use SAS to plot the two previous bar charts on the same bar chart to more effectively depict the similarities and differences of the two sexes easier. Using SAS, we can either plot the two bar charts are a stacked bar chart or a clustered bar chart:

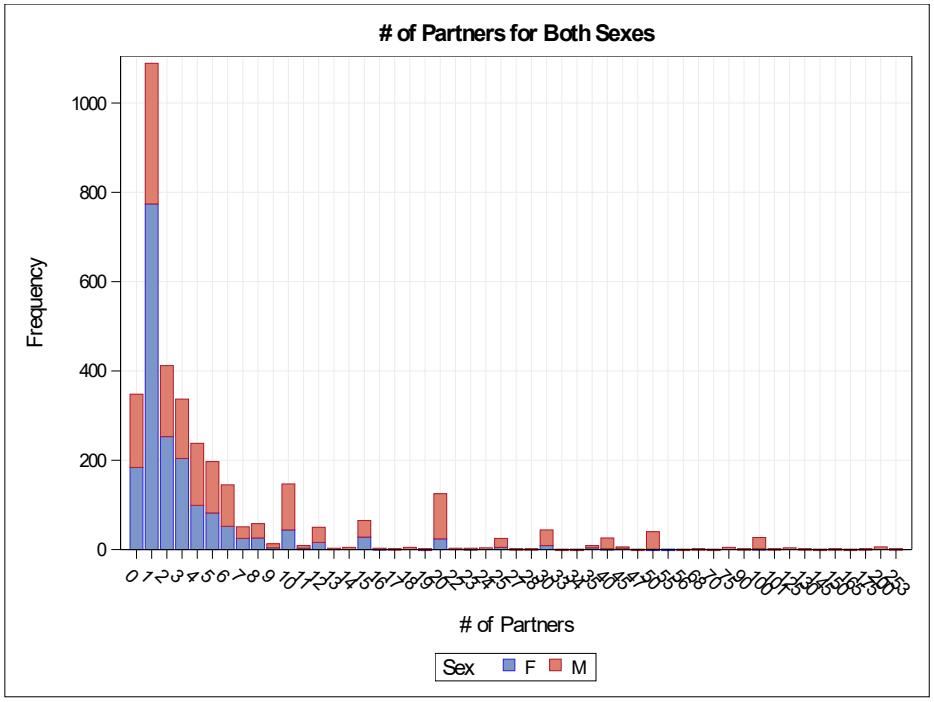

Figure 7 - Bar Chart of Frequencies for the # of Sexual Partners (Stacked)

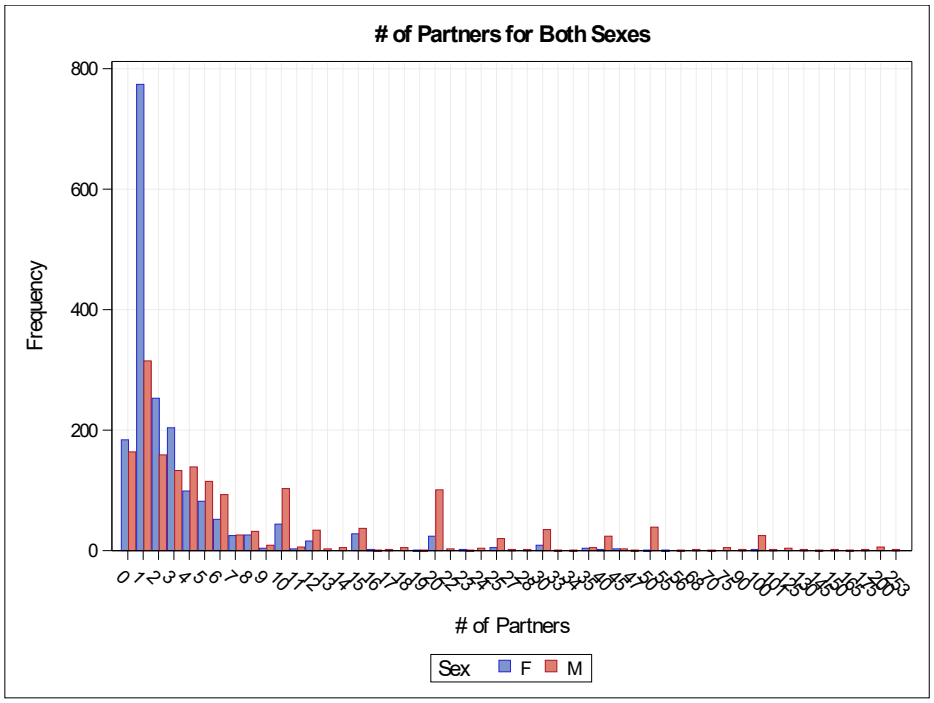

Figure 8 - Bar Chart of Frequencies for the # of Sexual Partners (Clustered)

Both graphs depict the same information, however, the clustered graph in my own opinion allows for a better comparison of the frequencies of the number of partners between the two sexes.

Looking at both graphs, the number of females who have only had one sexual partner is a lot larger than the males who have only had one sexual partner. We can also see that there are more females with 0-3 sexual partners than males. From 4 sexual partners onwards, Males always have more frequencies. As discovered before but made apparent again, the spread of male sexual partners is a lot larger than females.

CODE

```
/* DIRECT SAS TO FILE LOCATION */
FILENAME REFFILE
'/folders/myfolders/1/data/SA.csv'
;
/* IMPORT DATASET INTO SAS */
PROC IMPORT DATAFILE=REFFILE
DBMS=CSV
OUT=SA;
GETNAMES=YES;
RUN;
ODS SELECT VARIABLES;
/*PRODUCE THE CONTENTS TABLE OF
THE DATASET*/
PROC CONTENTS DATA=SA; RUN;
ODS SELECT DEFAULT;
/*change partners variable to
numerical value*/
DATA SA2; SET SA;
INT_PARTNERS = INPUT(PARTNERS,
BEST5.);
RUN;
/*PRODUCE THE CONTENTS TABLE OF
THE DATASET*/
ODS SELECT VARIABLES;
PROC CONTENTS DATA=SA2; RUN;
ODS SELECT DEFAULT;
/*PRODUCE THE STAT DESC OF THE
DATASET*/
PROC MEANS DATA=SA2;
VAR INT_PARTNERS;

RUN;
/*PRODUCE Bar chart of the average
number of partners per sex*/
PROC SGPLOT DATA=SA2;
VBAR SEX / RESPONSE = INT_PARTNERS
STAT=MEAN;
XAXIS GRID;
YAXIS GRID;
TITLE 'Average (Mean) # of
Partners For Each Sex';
LABEL INT_PARTNERS = '# of
Partners';
RUN;
/*PRODUCE new data set that only
contains males*/
DATA SA3;
   SET SA2;
   WHERE(SEX='M');
RUN;
/*PRODUCE bar chart of frequency
of # of partners for males*/
PROC SGPLOT DATA=SA3;
VBAR INT_PARTNERS;
XAXIS GRID;
YAXIS GRID;
TITLE '# of Partners for Males';
LABEL INT_PARTNERS = '# of
Partners';
RUN;
/*PRODUCE new data set that only
contains females*/
DATA SA4;
   SET SA2;
   WHERE(SEX='F');
RUN;
/*PRODUCE bar chart of frequency
of # of partners for females*/
PROC SGPLOT DATA=SA4;
  VBAR INT_PARTNERS;
  XAXIS GRID;
YAXIS GRID;
TITLE '# of Partners for Females';
LABEL INT_PARTNERS = '# of
Partners';
RUN;
/*PRODUCE grouped bar chart of
frequency of # of partners for
both sexes*/
PROC SGPLOT DATA=SA2;
  VBAR INT_PARTNERS / GROUP=SEX;
    XAXIS GRID;
YAXIS GRID;
TITLE '# of Partners for Both
Sexes';
LABEL INT_PARTNERS = '# of
Partners';
  RUN;
 /*PRODUCE clustered grouped bar
chart of frequency of # of
partners for both sexes*/

PROC SGPLOT DATA=SA2;
  VBAR INT_PARTNERS / GROUP=SEX
GROUPDISPLAY = CLUSTER;
    XAXIS GRID;
YAXIS GRID;
TITLE '# of Partners for Both Sex-
es';
```

CONCLUSION

Using the Statistical Analysis program SAS, along with its proprietary language, we created visual and numerical representations of two dataset, showing how powerful of a tool SAS can be.

Looking at both Datasets, we were able to derive more information by describing the data in several representation both numerical and visual. We were able to apply statistical operations such as averaging and different regression methods to gain a deeper understanding of the data along with the visual representation using scatter graphs and bar charts.

SOURCES

Giagos, D. and Shea, D. (2018). *Descriptive Statistics*. Manchester: Manchester Metropolitan University.

Matange, S. and Allison, R. (2018). *Stacked Bar Chart with Segment Labels*. [online] Graphically Speaking. Available at: https://blogs.sas.com/content/graphicallyspeaking/2013/09/20/stacked-bar-chart-with-segment-labels/ [Accessed 24 Oct. 2018].

Stat.purdue.edu. (2018). *Linear regression: SAS instruction*. [online] Available at: http://www.stat.purdue.edu/~tqin/system101/method/method_linear_sas.htm [Accessed 24 Oct. 2018].

Wicklin, R. (2018). *What is loess regression?*. [online] The DO Loop. Available at: https://blogs.sas.com/content/iml/2016/10/17/what-is-loess-regression.html [Accessed 24 Oct. 2018].

www.tutorialspoint.com. (2018). *SAS Bar Charts*. [online] Available at: https://www.tutorialspoint.com/sas/sas_bar_charts.htm [Accessed 24 Oct. 2018].

www.tutorialspoint.com. (2018). *SAS Scatter Plots*. [online] Available at: https://www.tutorialspoint.com/sas/sas_scatterplots.htm [Accessed 24 Oct. 2018].

YouTube. (2018). *Create a Simple Bar Chart Using SAS*. [online] Available at: https://www.youtube.com/watch?v=M-tCpWIF564 [Accessed 24 Oct. 2018].

YouTube. (2018). *Create Simple Scatter Plots with Two Variables Using SAS*. [online] Available at: https://www.youtube.com/watch?v=u2zC4HZJWO0 [Accessed 24 Oct. 2018].

YouTube. (2018). *SAS in 60 Seconds! - Converting Character to Numeric and Vice Versa*. [online] Available at: https://www.youtube.com/watch?v=SHOHsXFXDBo [Accessed 24 Oct. 2018].