**Visualising Data Final Assignment**

***Jake Bolger – C18395341***

**Find an appropriate dataset. There must be at least 8 attributes (in the final, merged, and tidied dataset) and 200 observations.  (4 marks)**

Dataset: youth-dataset-long-form.csv

**Description of my dataset and its origin:** The dataset that was used in this assignment was called the “National Longitudinal Survey of Youth 1979-2012”. This dataset is a project hat is based off a sample of American youth born between 1957 and 1964 on different aspects of life from the years 1979-2012. The dataset used in this assignment is a subset of this data. there are four main areas it focuses on, economic status, employment, education, and marriage. The dataset observes over 240,000 people. Whilst this dataset has lots of columns, the most important attributes for this assignment will be based off the gender, year of birth, and income.

**Link to dataset:** https://dasil.sites.grinnell.edu/downloadable-data/

**Tidy and / or enhance your dataset by merging. (4 marks)**

When tidying the dataset, the csv file was first read into a data frame. Then, the dataset was tidied by looking at the unique column and lengths of the attributes to see which ones were the best options. The columns that were selected for the new merged dataset are shown below:

**A screenshot of a computer

Description automatically generated with medium confidence**

The dataset was then written to a new csv file called “AmericanYouthLifeAspects” and read into a data frame. Although only the columns that were needed were being used, there were way too many observations amounting to almost 250,000. This would cause problems when making the graphs. To combat this issue, I created two new data frames which held a much smaller number of observations. One was based of the years 1979-1999, and the other was for 2000-2012. Both these new data frames had around 200-350 observations each.

Below are the data frames with the data that were used for a lot of the charts:

Graphical user interface, text, application

Description automatically generated

**Explore your data using charts and code.  (4 marks)**

**An explanation of what was discovered when you explored the data:**

When exploring the data, the first thing I looked at was the different time trends. Bar charts and line charts were used to do this. It was discovered that the average income has had a steady increase every year since 1979 for all genders. The maximum average income for the years 1979-1999 was around $30,000 whilst in the later years it increased to around $60,000. The female max income has also shown a rapid increase over the years whilst the male maximum income has shown a steadier increase.

Trend graph bar charts were created to look at the average income over the years. One for the years 1979-1999 and one for 2000-2012.

Chart, bar chart, histogram

Description automatically generated

Figure - Bar chart years 1979-1999

**Chart, bar chart

Description automatically generated**

Figure - Bar chart years 2000-2012

Next, line charts were created for the years 1979-2012. This time, the average maximum income was used instead of just the income.

Chart, line chart

Description automatically generated

Figure - Line chart Maximum Income 1979-99

Chart, line chart

Description automatically generated

Figure - Line chart Maximum income 2000-12

These line charts, combined bar charts were then created and combined using the max Income male and females grouped by year.

Chart, line chart

Description automatically generated

Figure - Combined line chart

Chart, bar chart

Description automatically generated

Figure - Combined Bar Chart

When exploring the data, the second thing I looked at was the different Comparison Charts. Histograms and density plots using single variable exploration, tree maps using simple comparison, and overlaid histograms for multi distribution comparisons were explored. It was discovered that there most people from the 1979-1999 data frame were born from 1957-60, and most people born from the 2000-2012 data frame were born from 1957-58.

It was then discovered that females have a higher density of lower income per year overall from the years 1979-2012.

Finally, it was discovered that most of the data represents people that were born in the US and that overall males that were born in the US had a higher income from the years 1979-1999.

Histogram charts and a density plot using single variable exploration were created to look at the amount of people born during the years 1957-64, and the max income based on sex. One for the years 1979-1999 and one for 2000-2012.

Chart, histogram

Description automatically generated

Figure - Single var Histogram

Chart, histogram

Description automatically generated

Figure - Single var histogram 2

Chart, histogram

Description automatically generated

Figure - Density plot Single var

Next, a tree map was created for the years 1979-1981 using simple comparison. This time, the income for both male and female was looked at based whether they were born on another country or born in the US.

Chart, treemap chart

Description automatically generated

Figure - Tree map income for gender based on country of birth

Finally, an overlaid histogram was created to explore multi distribution comparisons. For the income based on gender and income based on country of birth.

Graphical user interface

Description automatically generated

Figure – Multi-distribution Overlaid

Chart

Description automatically generated

Figure - Multi Country of birth

**Form and state your big idea. (2 marks)**

**BIG IDEA IS:**

The pay gap between men and women in recent years (2000+) is not as big as it used to be (1979-1999), the number of men and women with similar pay is increasing which should make women feel more equal and not feel like they are being discriminated against by getting paid a lot less than most men like before.

**Process of forming the big idea:**

1. Who is your audience?

The primary groups or individuals to whom you will be communicating.

* **Men**
* **Women**
* **Adults**
* **Young People**
* **Elderly**
* **All People**

1. If you had to narrow that down to a single person, who would it be?

* **People who still think there is huge inequality in pay between men and women.**

1. What does your audience care about?

* **If there is a pay gap between men and women.**
* **They care about if men are still getting paid a lot more than women.**

1. What action does your audience need to take?

* **Look at the data research provided.**
* **Realise that the pay gap between men and women isn’t that bad anymore and it’s getting better.**

1. What is at stake?

* **What are the benefits if your audience acts in the way you want them to?**
* **Women feeling like they are more equal**
* **Preventing the spread of misinformation on inequality.**

What are the risks if they do not?

* **Feel like they are inequal**

**Develop your story using 3 or more charts.  Charts should be relevant, effective, correct and should portray your story.  Your charts should be understandable by readers who have not read your background research.       (12 marks)**

**Your visualisation, with images of each of your charts:**

Encoded Scatterplots

These charts plot the data from two different time periods. One from 1979-1999 and the other from 2000-2012. They both show the income per year based on whether that person was male or female. As you can see the males had significantly more pay in most case. The highest income overall was also male, which was around $100,000, and the highest for the female was about the same as the average income of males. This shows a significant gap and pay between male and female.

**Chart, scatter chart

Description automatically generated**

Figure – Scatterplot

In this second plot, you can see that the income of male and females per year is a lot more even. For the most part the male and female population seem to earn close to the same amount. There are a couple of outliers over the years, some years a small proportion of the males have a significantly higher pay. However, this is also true for the female population, as you can see the highest recorded income was female, which was around $350,000 whilst the highest male was $280,000.

**Chart, line chart

Description automatically generated**

Figure - scatterplot

Encoded Density Plots

The charts below depict similar data as the scatter plots except they measure the density of male and females with a certain maximum income in two different time periods which are 1979-1999 and 2000-2012.

The first one for 1979-1999 shows that there is a much higher density of females that earn less than $60,000 a year compared to the males. Although there are a lot of males that earn that amount, they are spread out more even so that a large portion earn above $60,000.

Chart, histogram

Description automatically generated

Figure - Density Plot

For this period, (2000-2012), the graph shows that although the females still have a higher density of lower income, the males have a significantly closer density to the females. You can also see that there is a portion of females that earn large amount of income compared to the males.

This shows that although the pay isn’t completely equal, it is much more even than it was before.

**Chart

Description automatically generated**

Figure - Density Plot

Overlaying Histograms

For these graphs, histograms were created to shows the amount of people based on their gender who worked multiple jobs. As you can see there was more men that worked more jobs than women in 1979-1999. This could be a potential variable as to why they earned more.

Chart, line chart

Description automatically generated

Figure - Overlaying histogram

For this next histogram for 2000-2012, you can see that the number of jobs people work is more even based on gender. However, it still seems that men work slightly more.

**Chart, histogram

Description automatically generated**

Figure - Overlaying histogram

Stream Graphs

These stream graphs represent the mean income for the male and female population in the years 1979-1999. As you can see the male mean income is significantly higher than the females. However, in the next graph (2000-2012), the mean income is almost completely even which shows that the gap is nowhere near as bad as it was.

**Chart, histogram

Description automatically generated**

Figure - Stream Graph

**Chart, line chart

Description automatically generated**

Figure - Stream Graph

Small Multiples

These final graphs represent the maximum income grouped by gender based on the year someone was born. The first graph represents the years 1957-1960. This graph shows the year they were born has no impact on future oncome whether they are female or male. The same was said for the next graph for the years 1961-1964.

**Chart

Description automatically generated**

Figure - Small Multiples

**Chart

Description automatically generated**

Figure - Small Multiples

**\*\*Do at End\*\***

**Change column names to look better**

**Visualising text**

**Make charts perfect**