Problem Set 1: Visualization

Due: 5pm on Wednesday, January 31.

Student identifer: 5956

- Use this template to complete the problem set
- In Canvas, you will upload the PDF produced by your .qmd file
- Put your identifier above, not your name! We want anonymous grading to be possible

This problem set involves both data analysis and reading.

Data analysis

This problem set uses the data `lifeCourse.csv`.

```
library(tidyverse)
library(scales)
lifeCourse <- read_csv("https://info3370.github.io/data/lifeCourse.csv")
```

The data contain life course earnings profiles for four cohorts of American workers: those born in 1940, 1950, 1960, and 1970. Each row contains a summary of the annual earnings distribution for a particular birth cohort at a particular age, among the subgroup with a particular level of education. To prepare these data, we aggregated microdata from the Current Population Survey, provided through the Integrated Public Use Microdata Series.

The data contain five variables.

1. `quantity` is the metric by which the earnings distribution is summarized: 10th, 50th, or 90th percentile
2. `education` is the educational subgroup being summarized: College Degree, Less than College
3. `cohort` is the cohort (people with a given birth year) to which these data apply: 1940, 1950, 1960, 1970
4. `age` is the age at which earnings were measured: 30–45

5. `income` is the value for the given earnings percentile in the given subgroup. Income values are provided in 2022 dollars

1. Visualize (25 points)

Use `ggplot` to visualize these data. To denote the different trajectories,

- make your plot using `geom_point()` or `geom_line()`
- use the x-axis for `age`
- use the y-axis for `income`
- use `color` for `quantity`
- use `facet_grid` to make a panel of facets where each row is an education value and each column is a cohort value

You should prepare the graph as though you were going to publish it. Modify the axis titles so that a reader would know what is on the axis. Use appropriate capitalization in all labels. Try using the `label_dollar()` function from the `scales` package so that the y-axis uses dollar values.

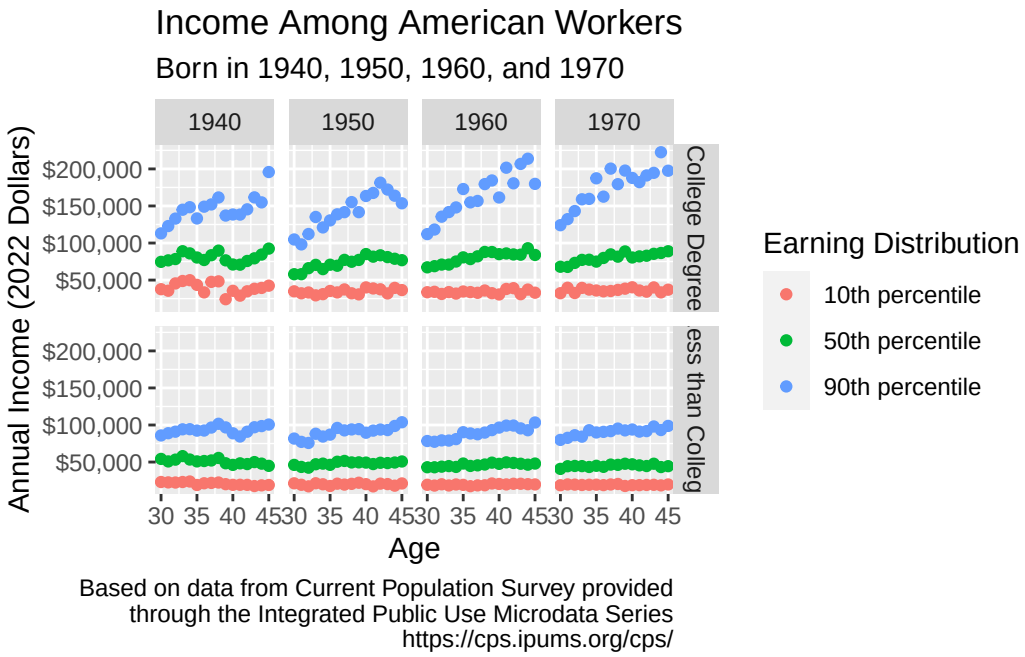
Your code should be well-formatted as defined by R4DS. In your produced PDF, no lines of code should run off the page.

Many different graphs can be equally correct. You will be evaluated by

- having publication-ready graph aesthetics
- code that follows style conventions

```
# your code goes here
lifeCourse |>
  ggplot(aes(x = age, y = income, color = quantity)) +
  geom_point() +
  facet_grid(row = vars(education), cols = vars(cohort)) +
  scale_x_continuous(name = "Age") +
  scale_y_continuous(
    name = "Annual Income (2022 Dollars)",
    label = label_dollar()
  ) +
  labs(
    title = "Income Among American Workers",
    subtitle = "Born in 1940, 1950, 1960, and 1970",
    caption = "Based on data from Current Population Survey provided
    through the Integrated Public Use Microdata Series
    https://cps.ipums.org/cps/"
```

```
) +
guides(color = guide_legend (title = "Earning Distribution"))
```



2. Interpret (10 points)

Write 2-3 sentences summarizing the trends that you see in the data.

2.1 (3 points). Focus on those born in 1970. For those with a college degree, how do the top and bottom of the income distribution change over the life course?

> For the American workers born in 1970 with a college degree, as age increases the difference in the top and the bottom of the income also increases– as demonstrated by the gap changing from approximately 90000 to 150000. For those within this group that earns a 10th percentile income, their annual income remains relatively the same as age increases. On the other hand, there is a clear increasing trend for annual income for those that earns a 90th percentile income as age increases, as demonstrated by the steeper positive slope.

2.2 (3 points). Focus on those born in 1970. How does the pattern differ for those without college degrees differ from your answer in 2.1?

As the age of those born in 1970 without a college degree increases, the difference between the top and the bottom of the income increases at a much slower rate as compared to those born in 1970 with a college degree– the gap between top and bottom only changed from 60000 to 80000. A starking difference is within the increase in income as age increases for those earning at the 90th percentile – as age increases for those without a college degree, the increase in annual income is very small compared to those with a college degree.

2.3 (4 points). How do the patterns you identified in 2.1 and 2.2 change from the 1940 to the 1970 cohort?

Similar to the 1970, the trend about wage increasing more for those with a college degree as compared to those without a degree as age increases still applies to the 1940, 1950, and 1960 cohort. However, the general rate of wage change is smaller for the other cohorts as compared to the 1970 cohort, as demonstrated by the wider fanning out of data points as we move from 1940 to 1970. In addition, for the cohort of 1940, there are periods where the 1940 cohort with college degrees earning at 10th percentile and those without college degrees earning at 90th percentile observe fluctuations, which are unusal for the other cohorts to observe large flunctuations.

3. Connect to reading (15 points)

Read p. 1–7 of following paper. Stop before the section "Analytic Framework for Decomposing Inequality."

Cheng, Siwei. 2021. The shifting life course patterns of wage inequality.. *Social Forces* 100(1):1–28.

Our data are not the same as Cheng's. But our analysis is able to reproduce many of her findings. Answer each question in two sentences or less.

Cheng discusses period trends, cohort trends, and age trends.

3.1 (3 points) Which dimension of your graph shows a cohort trend?

Since the graph is divided into cohorts, each column of the faceted panel shows a specific cohort and the cohort trend.

3.2 (3 points) Which dimension of your graph shows an age trend?

The x-axis of each faceted panel shows age trends.

3.3 (3 points) Cheng discusses education-based cumulative advantage. Describe how you see this in your graph.

In support of Cheng's argument about higher education leading to larger chances of upwards mobility, the graph shows the income and change in income from age 30 to age 45 being significantly greater for people with a college degree as opposed to those without, regardless of the earning percentile.

3.4 (3 points) Cheng discusses within-education trajectory heterogeneity. Describe how your graph shows heterogeneity of outcomes within educational categories.

For all cohorts with an college degree, the graph displays the high differences in income and wage inequality, which increases as age increases and demonstrates Cheng's claim about the fanning out of work trajectories as work specifics vary. Although the less obvious, there is a slight fanning out in income among the people without a college degree.

3.5 (3 points) Cheng discusses wage volatility: how wages rise and fall over time for a given person. Why is our data (the Current Population Survey) the wrong dataset to study wage volatility?

This dataset from the Current Population Survey is the wrong dataset to study wage volatility because this dataset only displays a summary of the annual earnings rather than individuals earnings.

Grad question: Model-based estimates

This question assumes familiarity with Ordinary Least Squares.

- For graduate students, this question is worth 20 points.
- For undergraduate students, this question is optional and worth 0 points.

The data contain nonparametric estimates that contain some noise: the data points provided partly reflect random variation because they are estimatd in a sample.

Model-based estimates reduce noise by pooling information across observations, at the cost of introducing assumptions. Fit an OLS model to the data using `age` as a numeric variable and `education`, `cohort`, and `quantity` as factor variables. Interact all of these predictors with each other.

```
fit <- lm(
  income ~ age * factor(cohort) * education * quantity,
  data = lifeCourse
)
```

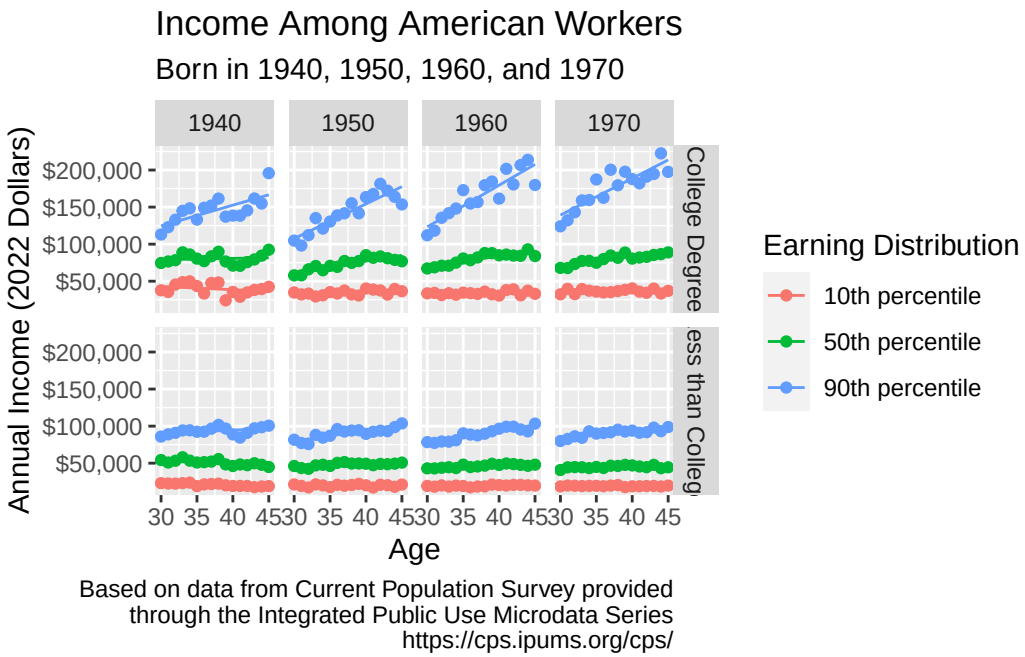
Effectively, this estimates a best-fit line through each set of points depicted in your original figure. For each observation, store a prediction from this model (see `predict()`).

Re-create your plot from (1) using

- `geom_point()` for the nonparametric estimates (as above)
- `geom_line()` for your model-based predicted values

```
predictions <- predict(object = fit)

lifeCourse |>
  ggplot(aes(x = age, y = income, color = quantity)) +
  geom_point() +
  facet_grid(row = vars(education), cols = vars(cohort)) +
  scale_x_continuous(name = "Age") +
  scale_y_continuous(
    name = "Annual Income (2022 Dollars)",
    label = label_dollar()
  ) +
  labs(
    title = "Income Among American Workers",
    subtitle = "Born in 1940, 1950, 1960, and 1970",
    caption = "Based on data from Current Population Survey provided
    through the Integrated Public Use Microdata Series
    https://cps.ipums.org/cps/"
  ) +
  guides(color = guide_legend (title = "Earning Distribution")) +
  geom_line(aes(y = predictions))
```



Computing environment

Leave this at the bottom of your file, and it will record information such as your operating system, R version, and package versions. This is helpful for resolving any differences in results across people.

```
sessionInfo()
```

```
R version 4.3.2 (2023-10-31)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Monterey 12.5.1

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;

locale:
[1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

time zone: America/New_York
tzcode source: internal
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] scales_1.3.0    lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
 [5] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.0
 [9] tibble_3.2.1    ggplot2_3.4.4   tidyverse_2.0.0

loaded via a namespace (and not attached):
 [1] bit_4.0.5         gtable_0.3.4      jsonlite_1.8.8    crayon_1.5.2
 [5] compiler_4.3.2    tidyselect_1.2.0  parallel_4.3.2    yaml_2.3.8
 [9] fastmap_1.1.1     R6_2.5.1          labeling_0.4.3    generics_0.1.3
[13] curl_5.2.0        knitr_1.45        munsell_0.5.0     pillar_1.9.0
[17] tzdb_0.4.0        rlang_1.1.3       utf8_1.2.4        stringi_1.8.3
[21] xfun_0.41         bit64_4.0.5       timechange_0.3.0  cli_3.6.2
[25] withr_3.0.0       magrittr_2.0.3    digest_0.6.34     grid_4.3.2
[29] vroom_1.6.5       rstudioapi_0.15.0 hms_1.1.3         lifecycle_1.0.4
[33] vctrs_0.6.5       evaluate_0.23     glue_1.7.0        farver_2.1.1
[37] fansi_1.0.6       colorspace_2.1-0  rmarkdown_2.25    tools_4.3.2
[41] pkgconfig_2.0.3   htmltools_0.5.7
```