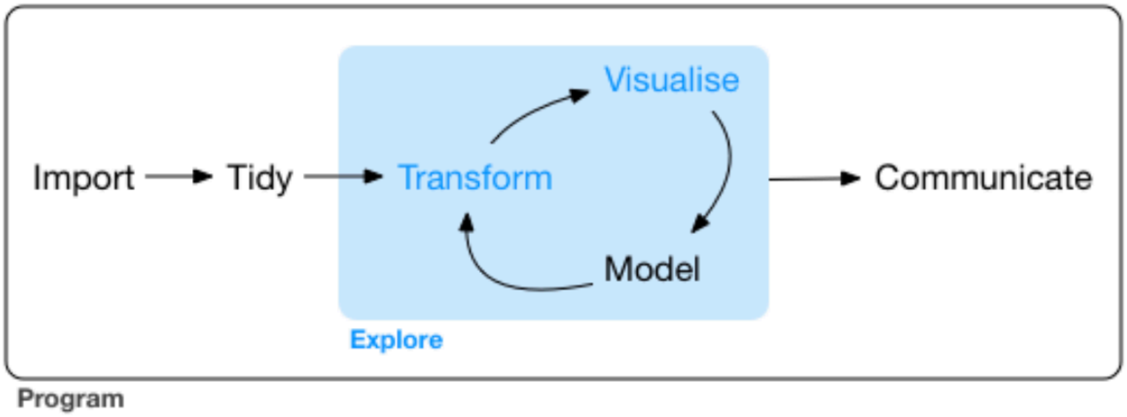
PhD Qualifying Examination

Spring 2020

CAP 5768 – Introduction to Data Science

Question 1 - [50 pts]

Use the diagram below[1] to answer the following questions:



(a) [15 pts] What processes are typically involved in the first two steps (*Import* and *Tidy*) and how much of the data scientist's time is usually spent in those steps?

It involves acquiring and cleaning up the data, storing it in a consistent form that matches the semantics of the dataset with the way it is stored.
It is estimated that 70-80% of the data scientist's time is spent doing this.

(b) [25 pts] The three steps indicated in the shaded blue area are often referred to as Exploratory Data Analysis (EDA). Explain in your own words the main processes involved in EDA and why they are so important in the context of a data scientist's work.

Transformation includes narrowing in on observations of interest (like all people in one city, or all data from the last year), creating new variables that are functions of existing variables (like computing speed from distance and time), and calculating a set of summary statistics (like counts or means). Together, tidying and transforming are often also called **data wrangling**, because getting your data in a form that's natural to work with often feels like a fight!

Once you have tidy data with the variables you need, there are two main engines of knowledge generation: visualization and modelling. These have complementary strengths and weaknesses so any real analysis will iterate between them many times.

[1] Figure credit: https://r4ds.had.co.nz/explore-intro.html

Visualization is a fundamentally human activity. A good visualization will show you things that you did not expect or raise new questions about the data. A good visualization might also hint that you're asking the wrong question, or you need to collect different data. Visualizations can surprise you, but don't scale particularly well because they require a human to interpret them.

Models are complementary tools to visualization. Once you have made your questions sufficiently precise, you can use a model to answer them. Models are a fundamentally mathematical or computational tool, so they generally scale well. Even when they don't, it's usually cheaper to buy more computers than it is to buy more brains! But every model makes assumptions, and by its very nature a model cannot question its own assumptions. That means a model cannot fundamentally surprise you.

(c) [10 pts] What processes are typically involved in the last step (*Communicate*) and how important is that step in the big scheme of things?

The last step of data science is **communication**, an absolutely critical part of any data analysis project. It doesn't matter how well your models and visualization have led you to understand the data unless you can also communicate your results to others.

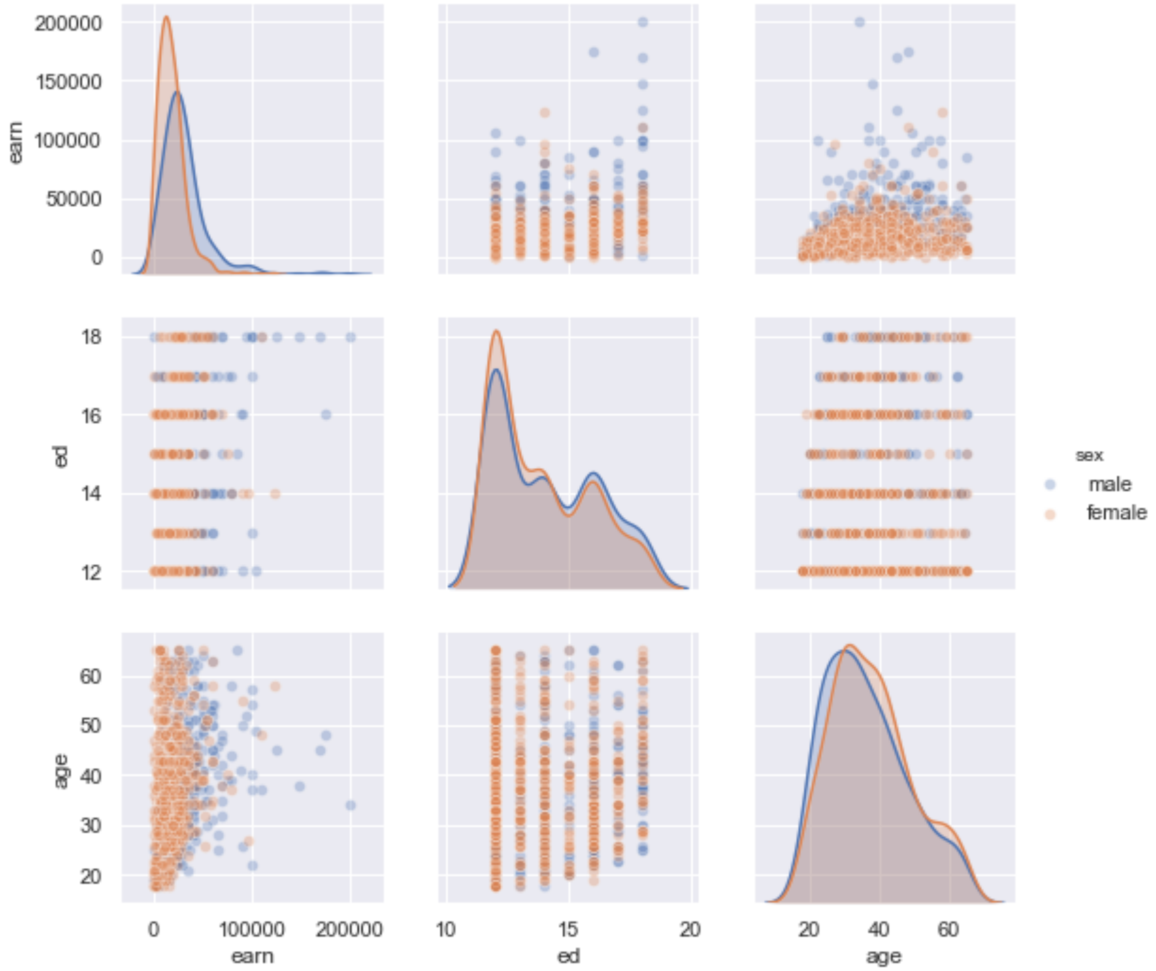
Question 2 - [35 pts]

Assume a small dataset with salaries from a fictitious company.
Each record (row) contains 6 (six) attributes (columns) as follows:

- `earn` is the annual salary.
- `height` is the height in inches.
- `sex` is a categorical value.
- `ed` is the number of years of education.
- `age` is the complete number of years of age.
- `race` is a categorical value.

Our objective is to check if there are signs of gender discrimination in the salaries.

Use the diagram below to answer the following questions regarding hypothesis testing.



(a) [15 pts] What does each of these 9 subplots represent and what type of preliminary information about (potentially unfair) pay gaps do they show?

From the pairplots we can spot a few pieces of information.

The text below refers to graphs by (*row*, *column*), for example (`earn`, `ed`) is the graph in the first row (`earn`), second column (`ed`).

- (`earn`, `earn`): the distribution shows that females earn less than males in general.
- (`earn`, `ed`): this scatter plot indicates that for the same education level, males earn more.
- (`earn`, `age`): this scatter plot indicates that for the same age, males earn more.
- (`ed`, `ed`): the distribution shows that females and males have about the same level of education.
- (`age`, `age`): the distribution shows that males and females have about the same age range.

Because the education and the age distributions (our proxy features) are the same for females and males, while the female salaries are lower, **we have an initial indication that there is gender discrimination.**

(b) [20 pts] How would you go about running a *rigorous hypothesis testing routine* on the data? Please be specific (state the null hypothesis, describe which tests you'd run, which calculations you'd perform, etc.)

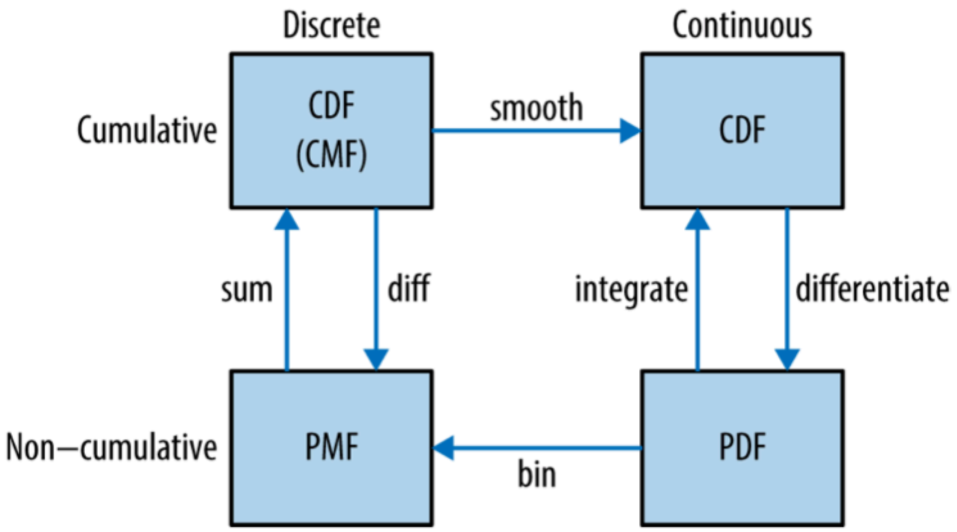
Null hypothesis: there is no gender discrimination.

Statistical test: Compute (Pearson's) Correlation coefficient and/or other metrics of correlation/covariance.

Check for statistical significance (i.e., compute p-value and ensure that it is < 0.005).

Question 3 - [15 pts]

Explain the diagram below[2] in your own words in the context of statistical distribution functions:



Note:
PDF = probability density function
PMF = probability mass function
CDF = cumulative distribution function
CMF = cumulative mass function

[2] Figure credit: https://greenteapress.com/wp/think-stats-2e/

- PMFs represent the probabilities for a discrete set of values.
  - To get from a PMF to a CDF, you add up the probability masses to get cumulative probabilities.
  - To get from a CDF back to a PMF, you compute differences in cumulative probabilities.
- A PDF is the derivative of a continuous CDF (i.e., a CDF is the integral of a PDF).
  - Remember that a PDF maps from values to probability densities; to get a probability, you have to integrate.
- To get from a discrete to a continuous distribution, you can perform various kinds of *smoothing*.
  - One form of smoothing is to assume that the data come from an analytic continuous distribution (like exponential or normal) and to estimate the parameters of that distribution.
  - Another option is kernel density estimation.
- The opposite of smoothing is *discretizing*, or *quantizing*.
  - If you evaluate a PDF at discrete points, you can generate a PMF that is an approximation of the PDF.
  - You can get a better approximation using numerical integration.
- To distinguish between continuous and discrete CDFs, one could call a discrete CDF “cumulative mass function” (CMF).