Refers to 1994 data:

DAY 1:

Saturday 2/1

Setup environment from 9AM to 12AM

Analyze metadata from 3PM to 4PM

Exploratory Data Analysis: 4PM to 6PM

Summarize daily actions 6PM to 6:30PM

# SETUP

I am a big fun of R’s elegance and grace. And I am aware of the effectiveness of Python. After the 2020 rstudio conf in San Francisco, I decide this is a good chance to give reticulate a try.

Therefore, I will setup two virtual environment: one for Python (Conda) and one for R (RStudio).

Next I will work on the project as a rmarkdown combining Python and R.

In order to simplify visual prototyping I might give Power BI a chance.

I was not able to use Anaconda because RStudio is available only up to version 1.1. Instead, RStudio 1.2 has enhancements for using Python. <https://blog.rstudio.com/2019/04/30/rstudio-1-2-release/>

Next step I needed a package manager in order to mange in RStudio virtual environments for R and Python. Seems like renv is the suitable candidate <https://blog.rstudio.com/2019/11/06/renv-project-environments-for-r/>

Therefore, I checked out the documentation for RStudio and reticulate <https://docs.rstudio.com/tutorials/user/using-python-with-rstudio-and-reticulate/>

Versioning  
Nevertheless I created a new repository on github and synched my local repository to work on RStudio. This question <https://stackoverflow.com/questions/34565238/where-does-github-desktop-install-command-line-version-of-git> helped me find the git.exe executable installed in the github desktop distribution.

Also this was useful to setup RStudio and git. https://cfss.uchicago.edu/setup/git-with-rstudio/

I decided to inspire the structure of the project according to this <https://drivendata.github.io/cookiecutter-data-science/>

Considering important aspects is reproducibility, I will create virtual environment. Ideally, at the end of the project I would create a Docker container ready-to-use.

# Data Source

Ideally, I would have preferred not to manually download the data from the zip file sent by Marie, instead download it directly from AWS S3. Additionally, I would have preferred to keep all my data in AWS S3.

However, I added a .gitignore in order not to version changes in data using github. Instead, data must be permanent such that once created, a dataset is never overwritten. Instead, a different dataset is created upon transformation.

# Environment

The project is accomplished on my local computer. An alternative approach would have been to entirely build it using Docker container.

I created a conda virtual environment for Python.

I created a project library for R using renv.

## Dependency manager

# METADATA

I reorganized the metadata information in a simple table like the following using Excel:

Code| description|data\_type|possible\_values

Once organized the environment, I started preparing the metadata for two reasons:

1. Column headers are missing in the data files
2. Metadata file does not provide a clear mapping of column name and column position in the data file

I proceeded the metadata preparation in Excel because the columns were less than 50 so it was reasonable to proceed in Excel. The process was by exclusion: first assigned the column name to the field (e.g. Married will be marital status).

Then I remained with the field in position # 2 #3 not clear.

During this process I observed that in the 18 records I used for understanding the column names, nominal categorical variables might have required normalization through remapping.

In the end I observed that all variables related to income – except the target – were removed from the training set therefore this allowed me to solve doubts about remaining column positions.

The output of the process is “/docs/metadata/census-income-metadata-prepared.v00.csv”

# Workflow

I want to define a first iteration of EDA analyzing variables and their distribution split by target value.

The outcome I expect is:

* Identify promising variables
* Identify variables that might need treatment
  + Outliers
  + Remapping of nominal variables

Test set will remain not used.

All analysis are applied on the training set and defined replicable so they will be easily applied also to the test set.

# Exploratory Data Analysis

As a first step, I import the data.

An interesting point to investigate is the instance weight. I would like to understand whether it refers to training or also to test.

## Validation

# Remapping Categorical Variables

Considering the high level of remapping for normalizing the data, it will be important to deploy a mechanism to monitor when appears data outside the distribution.

# External data:

Pull external data for income about the different countries. However, many have missing values so it was not very useful.

Instead, it is useful to understand the median salary (if not available then also the average) for 94 and 95.

Ideally, we would retrieve this information for each state such that we can assign a priori probability to each individual.

<https://catalog.data.gov/dataset?organization_type=State+Government&tags=income>

http://kdd.ics.uci.edu

https://www.google.it/search?client=opera&hs=EeI&ei=55g1XsamK4WYkwX8raOgDw&q=median+personal+income+1995+U.s.&oq=median+personal+income+1995+U.s.&gs\_l=psy-ab.3...17404.18861..18898...0.3..0.105.1275.14j1......0....1..gws-wiz.......0i71.KR1dA2xoyY0&ved=0ahUKEwiGtpaj1bDnAhUFzKQKHfzWCPQQ4dUDCAo&uact=5

# Objective:

<https://archive.ics.uci.edu/ml/datasets/Census-Income+%28KDD%29>

https://kdd.ics.uci.edu/databases/census-income/census-income.data.html

Predict total personal income level as a dichotomous variable binned at the $50K

# Next steps

* Review code for production
  + E.g. Build in Dataiku
  + Connect to AWS data source
* Test on 2019 Census
  + This might require understanding whether attributes are the same
* Understand changes over time of different variables