Assignment 3

The Center for Disease Control and Prevention (CDC) is a federal public health and safety agency of the United States. Among other things, the CDC researches infectious diseases (like the flu) and collects public health data. The CDC will let the public explore their death dataset, which records cause-of-death of everyone deceased in the United States. The CDC's goal is to raise public awareness of health, so the dataset will only include deaths by natural causes or accidents, to help current residents of the country take better precautions about their own health. You can find the dataset on Kaggle:
https://www.kaggle.com/cdc/mortality
You can find some more information on how to interpret the data here:
https://www.kaggle.com/sohier/mortality-data-format-v2-tutorial/data

> You can download the dataset and work on Google Colab or locally, but it's also possible to create a notebook on Kaggle directly (using the option New Notebook). Make sure to regularly save/download your work, since session times are limited on Kaggle.

The data

The data 2015_data.csv contains information on United States residents that died from natural causes or accidents in 2015. Each row is a person. The accompanying file 2015_codes.json explains the available columns in the dataset.

Some important columns are highlighted below:

detail_age **detail_age_type**	age in years Only when detail_age_type == 1 (= years)
sex	sex, restricted to F (female) or M (male), at time of death
race	Including: ['White', 'Black', 'Korean', 'Vietnamese', 'Indian', 'Native American', 'Hawaiian', 'Chinese', 'Japanese', 'other Asian or Pacific Islander', 'Filipino', 'Samoan', 'Guamanian']

hispanic_origin	Specific subclassification for hispanics.
education_2003_revision	1 ... 8th grade or less 2 ... 9 - 12th grade, no diploma 3 ... high school graduate or GED completed 4 ... some college credit, but no degree 5 ... Associate degree 6 ... Bachelor’s degree 7 ... Master’s degree 8 ... Doctorate or professional degree 9 ... Unknown
month_of_death	numerical value of month
day_of_week_of_death	1 ... Sunday 2 ... Monday 3 ... Tuesday 4 ... Wednesday 5 ... Thursday 6 ... Friday 7 ... Saturday 9 ... Unknown
manner_of_death	E.g. ‘Natural Causes’ or ‘Accident’

marital_status	S ... Never married or Single M ... Married W ... Widowed D ... Divorced U ... Unknown
icd_code_10th_revision **39_cause_recode**	The exact cause of death as standardized by international medical classification codes. See https://icd.codes/icd10cm for reference The recode columns group these codes into a number of groups (39 in the case of 39_cause_recode).
record_condition_{1-20}	Multiple cause of death data. A combination of 1 or more ICD-codes.

Part 1: Explore & Visualize

[If you've never visualized in Python, Pandas has some visualization support. https://pandas.pydata.org/pandas-docs/stable/user_guide/visualization.html You can also use Matplotlib https://matplotlib.org/gallery/index.html]

The goal of these tasks are to have you experiment with different ways of visualizing the data.

Show your work in code as well as your final visualizations in a notebook. Label each question using markdown in the notebook and include answers to all questions.

Submit part 1 as a notebook assignment3-part1.ipynb

A. Create a histogram of **death counts by age.**

   Hint: one axis should be sorted continuous range of age 0 up to the age of the oldest person in the dataset

B. Create a histogram of deaths caused by **the flu by age**.

When shown data or an ML model, humans tend to have *confirmation bias*, meaning that they tend to believe that whatever the data or model says *is what they really thought all along*. Ever broke up with a significant other and your friends tell you “I told you so”? This is confirmation bias. With Bayesian reasoning, we can take into account a viewer’s *prior reasonable guess* before they see data. This is a good technique to help users reflect on how the data might *conflict* with “what they thought all along”.

C. Write down what you believe (before looking at the data. Just guess!) is the relationship between **age** and death by **Motor Vehicle Accident**. Do you expect the risk of death by car accident to be the same across all ages or higher in certain age ranges? Why?
D. Write down what you believe (before looking at the data. Just guess!) is the relationship between **age** and death by **Drug Use**. Do you expect the risk of death by drug use to be the same across all ages or higher in certain age ranges? Why?

E. Create a histogram of **death** by **Motor Vehicle Accident** by **age**. Create a histogram of **death** by **Drug Use** by **age**.
F. Compare your prior guess in C and D to the histograms in E. What did you learn from the histograms? Are there parts of your prior guess that were confirmed by the histograms? Are there parts of your prior guess that were wrong or different than you expected?

When users see different possibilities separately in a data or ML system, there’s a bias towards thinking *all possibilities are equally likely*, when really some options are more or less probable in real life. E.g., While a headache could be caused by fall allergies or by brain cancer, the likelihood of fall allergies is far higher in real life than brain cancer.

G. Create a visualization of your choice, where you overlay 4 different causes of death (your pick) by age the same plot. Design this visualization however you wish. Justify your design by writing a few sentences about how your visualization will help users compare the 4 different death risks by age. Talk about encoding choices such as: plot type, use of size, color, and axes labels.

Part 2: Designing Personal Predictions

The goal of Part 2 is to start designing an interactive interface, where a user that comes to the CDC visualization can put in their own information, and see the most common causes of death for their attributes (like age, gender, and so on)

Show your work in code as well as your final visualizations in a notebook. Label each question using markdown in the notebook and include answers to all questions.

Hint: To add some minimal interactivity with minimal effort, consider using Jupyter Notebook Widgets: https://ipywidgets.readthedocs.io/en/latest/examples/Widget%20List.html

Hint 2: It's actually rather easy to make it fully interactive. see https://towardsdatascience.com/interactive-controls-for-jupyter-notebooks-f5c94829aee6

also, https://ipywidgets.readthedocs.io/en/latest/examples/Using%20Interact.html

Submit part 2 as a notebook assignment3-part2.ipynb

A. Design for personas. For each of the fictional users given, create a single visualization that shows the most likely cause of death for that user. To experiment with design choices, make each user/visualization pair a *different visualization that represents different design choices* (e.g. you could try a different plot type for some users).

    - **Miles** is a young black male college student. He is 20 years old and lives a healthy lifestyle. He doesn't smoke or use drugs, but does occasionally drink alcohol to the extent it influences his behavior.
    - **Jonas** is a 72 year old man, immigrated from Germany to the United States in his thirties, and has a highschool level education. His wife passed away last year.
    - **Alma** is a 36 year old woman with two kids. She is hispanic and co-parents her kids with a long term romantic partner but does not believe in marriage. She has an accounting degree.

B. Which visualization from A do you think is the most successful design? What visualization techniques did you use?

C. Given your visualizations in A, what would be *good* questions for a user to ask a personalized visualization from this dataset? What would be some *bad* questions ie. questions that a personalized visualization (with this dataset alone) cannot answer?

D. If users like those in A visit the interactive tool on the CDC website, what information (e.g. age or race) would you have them put in to show the most relevant death visualization and why?

E. For each column in the dataset, describe how you would bin the data for a good user experience and why. For instance, ‘age’ can be not-binned (exact numbers) or binned into ‘child’, ‘young adult’, ‘adult’, ‘elder’... or many other bin choices.