University of Science and Technology
Communications & Information Engineering
CIE 457: Statistical inference and data analysis

Final Course Project:
Analyzing U.S. COVID-19 Data

In this project we will use the publicly available COVID-19 Case Surveillance data offered by The Center for Disease Control and Prevention (CDC) and the US Census Bureau to analyze the patterns of COVID-19 spread and management in the US across time, regions, and demographics.

Here is where you can download the dataset samples we'll use for this project:

COVID-19 Case Surveillance:
- DATA.
- Official website.
- Codebook.

Household Pulse Survey:
- DATA.
- Official website.
- Codebook.

> PLEASE NOTE: you will need to carefully read the datasets' **official websites** to understand how the data was collected, what it represents, and any other limitations you might need to be aware of. In addition, you'll need to carefully study the **codebooks** to understand the data format, structure, value ranges, interpretation, column names, etc.

In the next sections, you get to decide which data (or combinations of data) to use in each analysis task.

PART 1: Exploratory Analysis: [20%]

All analysis points must be accompanied by your commentary and at least one appropriate visualization.

Use the appropriate statistics and plots to investigate the following:

1. The total number of hospitalizations versus deaths from COVID-19 over the entire US per month-year timestamp.
2. The average rates of COVID-related deaths relative to patient **demographics**[1].
3. The rates of COVID-related hospitalization and death with age (across age groups).
4. Average rate of COVID-related hospitalization and death per state over the entire study period.
5. The relationship between age, pre-existing medical conditions and/or risk behaviors, and rate of admittance to the ICU.
6. The rate of expected employment loss due to COVID-19 and sector of employment.
7. The rate of expected employment loss due to COVID-19 relative to responders demographics.
8. The rate of expected employment loss due to COVID-19 for the top 10 states with the highest rate of COVID hospitalization.
9. The relationship between household income and the rate of delayed/ OR unobtained medical treatment (Due to COVID or otherwise).
10. The relationship between COVID-19 symptom manifestation and age group.

[1] **demographics: age, sex, race.**

PART 2: Answering Questions: [20%]

All analysis points must be accompanied by your commentary and at least one appropriate visualization.

2.1 Use the appropriate statistics and plots to answer the following questions:

1. Are hospitalized patients with underlying medical conditions and/or risk behaviors more likely to die from COVID-19?
2. Who are the people (the demographic segment) that appear to be most at risk of death due to COVID-19? Who is the least at risk?
3. What percent of patients who have reported exposure to any kind of travel / or congregation within the 14 days prior to illness onset end up hospitalized? What percent of those go on to be hospitalized?
4. Are Asymptomatic COVID patients less likely to be hospitalized? Are they less likely to die from their illness?
5. Which state is associated with the highest percentage of Economic Impact (stimulus) payments among survey respondents?

2.2 Come up with 5 more bivariate/multivariate analysis questions and similarly answer each with appropriate visuals and commentary.

PART 3: Hypothesis Testing: [15%]

Claim: "There is a strong association between probability of death due to COVID-19 and patient demographics"

3.1 Formulate a hypothesis test to assess the validity of this claim given the available data:

- State the test you will use and justify your choice.
- Clearly state the hypotheses.
- Conduct the test and report the result.
- Make a conclusion as to the validity of the claim, assume a significance level of 0.05.

3.2 Come up with your own claim from the available data and conduct a hypothesis test for it following in the same steps.

PART 4: Regression Analysis: [15%]

Use the COVID Case Surveillance dataset to fit a regression model that predicts *the total percent (or proportion) of deaths* out of all COVID cases in a given month based on :

- **Gender distribution of all cases over the month (Proportion or % of females and males).**
- **Age distribution of all cases over the month (Proportion or % of each age group).**
- **Proportion (or %) of all cases over the month that end up in the ICU.**
- **Proportion (or %) of all cases over the month that end up hospitalized.**

1. Report your model's coefficients and p-values.
2. Which of these variables are good predictors of the variabilities in the target? Which are bad ones?
3. Are any of these predictors correlated with each other?
4. Experiment with different ways to improve the fit and interpretability of the model. For example:
   - Add or remove the intercept.
   - Introduce higher order terms.
   - Remove outliers if any.

PART 5: Bonus Task:[10%]

Train a machine/deep learning classifier to predict the likelihood of death due to COVID-19 using any/all of the relevant attributes in the COVID case surveillance dataset.

PART 6: Documentation:[20%]

You are required to submit two documents:

1- <u>A business report:</u> containing all the analysis results from parts 1 to 5 along with the associated visualizations and comments. You should also include an introduction and a conclusion explaining the significance and potential limitations of your findings.

2- <u>Technical documentation:</u> Should explain:

- The structure of your project: file hierarchy, general flow, etcetera.
- A description of all functions and their usage.
- The steps you followed for data collection/cleaning, and all subsequent analysis requirements.
- The challenges/limitations/assumptions involved in any step.

For the technical documentation we will accept any of the following:

- A simple readme.md file.
- A pdf/word report.
- Or just markdown documentation within your notebook.

PART 7: Presentation:[10%]

For the final presentation you should prepare slides summarizing all your findings with visualizations and brief comments. You'll be graded on the quality of your visual aids and your ability to present your findings in a brief but compelling manner.