DATA-101
Exam I Take-Home

Exam Instructions

1. **Show all your work** and write complete and coherent answers.
2. Show all of the steps that you used to get your final answer. If you do not show your work I can not give partial credit in the case of incorrect answers.
3. **Please strive for clarity and organization.**
4. **You are not allowed to discuss any of the exercises in Exam 1 take-home with others. Identical submissions will receive a 0 as a grade in Exam 1 take-home.**
5. **Late submission will not be accepted, regardless of the circumstances.**

Take the time to carefully read all the questions on the exam. GOOD LUCK!

1. Consider the `insurance.csv` datafile. This file contains some basic demographic information related to an insurance company in the USA. **In Python**, answer the following:

   (a) (3 points) Using the pandas library, read the csv file and create a data-frame called `insurance`.

   (b) (3 points) Using the appropriate Python commands, report the number of observations and variables.

   (c) (3 points) Using the appropriate Python commands, report the number of females and males in the `insurance` data-frame.

   (d) (4 points) Create the histogram of `age`. Comment on the plot.

   (e) (5 points) A common hypothesis is that smokers insurance charges is higher than non-smokers. Compute the average charges for smokers and non-smokers. Does this calculation agree with the initial hypothesis? Explain. (Hint: you might consider using `groupby()`)

   (f) (5 points) A common hypothesis is that smokers insurance charges is higher than non-smokers. Compute the median charges for smokers and non-smokers. Does this calculation agree with the initial hypothesis? Explain. (Hint: you might consider using `groupby()`)

   (g) (5 points) Using the results from parts (e) and (f), what can you conclude about the shape of the distributions of charges of smokers and non-smokers? (Hint: compare the mean and median)

2. Consider the Wisconsin breast cancer diagnostic data set from the UCI Machine Learning Repository (`for more info click here`). The Wisconsin breast cancer data set contains information on 569 biopsies, each with 32 features. One feature is an id number, another is the cancer diagnosis, "`M`" to indicate malignant or "`B`" to indicate benign. The other 30 numeric measurements comprise the mean, standard error, and worst (that is, largest) value for 10 different characteristics of the digitized cell nuclei. Notice that all these features are related to the shape and size of the cell nuclei. **In Python**, answer the following:

   (a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `breast_cancer`.

   (b) (3 points) Using the appropriate commands, report the number of `B` and `M` diagnoses.

   (c) (5 points) Using the appropriate commands, report the summary statistics of `radius_mean` and `area_mean`. Comment on the shape of the distributions of `radius_mean` and `area_mean`. (Hint: compare the mean and median)

   (d) (8 points) The z-score standardization is a numerical measurement used in statistics of a value's relationship to the mean (average) of a group of values, measured in terms of standard deviations from the mean. The z-score standardization formula is given by

   $$z = \frac{X - \overline{X}}{s}$$

   where $\overline{X}$ and s represent the mean and standard deviation, respectively. If a z-score is 0, it indicates that the data point's score is identical to the mean score. A z-score of 1.0 would indicate a value that is one standard deviation from the mean.

z-scores may be positive or negative, with a positive value indicating the score is above the mean and a negative score indicating it is below the mean. Create two variables in the `breast_cancer` data-frame: one called `z_radius_mean` and another one called `z_area_mean` that represent the z-score standardizations of `radius_mean` and `area_mean`, respectively.

(e) (6 points) A typical application of the z-score is to flag out potential outliers in a data set. The rule is that any observation with a z-score greater than 3 or less than -3 is flagged as potential outlier. Report the five-number summary statistics of `z_radius_mean` and `z_area_mean`. Are the any outliers in those two variables? Explain. Be specific.

3. What is Data Science? Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data. A data scientist is someone who creates programming code and combines it with statistical knowledge to create insights from data. Consider the `data_science_job_salaries.csv` data file. This data file contains the following information:

   - `work_year`: The year the salary was paid.
   - `experience_level`: The experience level in the job during the year.
   - `employment_type`: The type of employment for the role.
   - `job_title`: The role worked in during the year.
   - `salary`: The total gross salary amount paid.
   - `salary_currency`: The currency of the salary paid as an ISO 4217 currency code.
   - `salaryinusd`: The salary in USD.
   - `employee_residence`: Employee's primary country of residence in during the work year as an ISO 3166 country code.
   - `remote_ratio`: The overall amount of work done remotely.
   - `company_location`: The country of the employer's main office or contracting branch.
   - `company_size`: The median number of people that worked for the company during the year.

   In Python, answer the following:

   (a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `ds_salary`.

   (b) (3 points) Using the appropriate commands, report the top three job title labels in the `job_title` variable.

   (c) (5 points) Using the appropriate commands, report the top five job titles with the highest average salaries (in USD). *Hint:* you might consider using `groupby()`.

   (d) (8 points) A common hypothesis these days is that data science employees, who come to the office regularly, make more money. Using the appropriate commands, compute the average salary (in USD) based on `remote_ratio` for US employees. Does this calculation agree with the initial hypothesis? Explain.