Project #1: Data analysis with R

Project Description and Requirements:
R is an open-source (that means, it is FREE! yes FREE!) software. R is widely used globally by Statisticians, Data Scientists, Data Analysts, and Data Engineers in major Tech Companies like Amazon, Google, and Microsoft. Given a large dataset with hundreds (or millions or billions) of observations, it is extremely difficult (or if not, almost impossible) to manually perform descriptive statistics of the dataset. Hence the need for statistical software to manipulate large datasets. This requires a basic knowledge of the chosen software, which is the main purpose of this project. The only requirement for this project is an interest in learning. Basic familiarity with computers is helpful. No prior coding or R or Data Science experience is needed. Students will need to download and Install R. The mentor will help students with all of that. Helpful but not required: experience/background in a scientific or engineering discipline. This project will help students gain a foundational understanding of a subject or tool, as well as enable students to develop job-relevant skills with hands-on projects.

Project Goals:

1. Installation of R.
2. Data importing and loading dataset into R.
3. Perform descriptive statistics using R.
4. Working with data types and variables.
5. Vectors, Lists, matrices, and data frames.
6. Logical statements such as if, for, while loops, iterations and Repeats.
7. Functions, data plotting and visualization.
8. Introduction to basic statistical functions and packages.

Project Tools: R software (https://www.r-project.org/). It is FREE!

Project Meeting Format, Schedule, and Timeline: This project may be completed between 1-month to 3-months, depending on student's background and/or availability. The student meets weekly face-to-face and/or virtually with mentor for 1-hour to 3-hours, as determined by need and availability.

Project #2: Identifying and dealing with outliers in a dataset.

Project Description and Requirements:

Outliers are defined to be extreme values in a dataset. When dealing with datasets, it is important to investigate the presence of outliers. Why is important to check for the presence of outliers in a dataset? In fact this matters because outliers can have a big impact on statistical analyses and may skew the results of a hypothesis test and/or other inferential statistics, if the outliers are inaccurate and/or they are not processed correctly. Outliers can also negatively impact the statistical analysis and result, making it hard to detect a true effect if there is one. We may sometimes need to remove and/or cap outliers from the dataset before doing any data analysis. Graphical displays such as histograms and boxplots may be used to visually detect outliers. Both displays can also show the researcher how the data is distributed. This project will help students gain a foundational understanding of a subject or tool, as well as enable students to develop job-relevant skills with hands-on projects.

Project Goals:

1. Investigate outliers using statistical methods.
2. Find outliers and view the data distribution using a histogram.
3. Find outliers in data using a box plot.
4. Drop and/or cap outliers.
5. Find multivariate outliers using a scatter plot.

Project Tools: The student is free to use any of the following statistical software as a data analysis tool: Python, R, SAS, Minitab, MATLAB, and Microsoft Excel.

Project Meeting Format, Schedule, and Timeline: This project may be completed between 1-month to 3-months, depending on student's background and/or availability. The student meets weekly face-to-face and/or virtually with mentor for 1-hour to 3-hours, as determined by need and availability.

Project #3: Data analysis fundamentals with Python

Project Description and Requirements:
Python is an open-source (that means, it is FREE! yes FREE!) software. Python is widely used globally by Statisticians, Data Scientists, Data Analysts, and Data Engineers in major Tech Companies like Amazon, Google, and Microsoft. Given a large dataset with hundreds (or millions or billions) of observations, it is extremely difficult (or if not, almost impossible) to manually perform descriptive statistics of the dataset. Hence the need for statistical software to manipulate large datasets. This requires a basic knowledge of the chosen software, which is the main purpose of this project. The only requirement for this project is an interest in learning. Basic familiarity with computers is helpful. No prior coding or Python or Data Science experience is needed. Students will need to download and Install python. The mentor will help students with all of that. Helpful but not required: experience/background in a scientific or engineering discipline. This project will help students gain a foundational understanding of a subject or tool, as well as enable students to develop job-relevant skills with hands-on projects.

Project Goals:
1. Installation of Python.
2. Data importing and loading dataset into Python.
3. Perform descriptive statistics using Python.
4. Working with data types and variables.
5. Vectors, Lists, matrices, and data frames.
6. Logical statements such as if, for, while loops, iterations, and Repeats.
7. Functions, data plotting and visualization.
8. Introduction to basic statistical functions and packages.

Project Tools: Python software (https://www.python.org/). It is FREE!

Project Meeting Format, Schedule, and Timeline: This project may be completed between 1-month to 3-months, depending on student's background and/or availability. The student meets weekly face-to-face and/or virtually with mentor for 1-hour to 3-hours, as determined by need and availability.