**Design Your Own Learning Homework**

**Goal:** Developing your data analysis skills, including data collection, cleaning, analysis, and visualization.

**Deadline:** End of second semester (May 15, 2023)

**Schedule of learning:**

**Week 1-2:** Data Analysis Overview:

Learn the basics of data analysis, covering ideas like data types, data structures, and data collecting.

Learn the fundamentals of Microsoft Excel and its data analysis tools, such as pivot tables, charts, and graphs. Python and Jupyter notebook installation

Materials: class lectures, YouTube videos

**Week 3-4:** Analysis of Data with Python and Excel.

Learn the fundamentals of Python and its libraries for data analysis, such as NumPy, Pandas, and Matplotlib, Overview of NumPy arrays, Python data types, construction, array creation functions.

**Resource:** DataCamp assignments Data analysis in excel, class lectures, review some materials from previous coursework TECH FOUNDATIONS CIS, code with mosh.

**Application:** Use of Excel to clean a real-world dataset. From a given dataset about 41 students response regarding various aspects, various kind of analysis using excel and find some outputs. Analysis includes cleaning text data, doing calculations using VLOOKUP, finding maximum, minimum, and average number, etc.

Use of python for data visualization: Created 2D array with 8 by 6 dimension, created array slices marked by different colors, created Boolean mask to identify all negative numbers of the given dataset and replace the negative numbers with zero, using Jupyter notebook.

**Week 5–6:** Introduction to Statistical Analysis: Learn the fundamentals of statistical theory, including the mean, median, mode, standard deviation, and hypothesis testing.

Revise course materials from last semester course work Quantitative analysis.

Data visualization: Learn the fundamentals of data visualization using pandas and how to read pandas.

**Resources:** Class lectures, books, online video tutorials.

**Application:** Cleaning and prepping the dataset in Python and Excel before analyzing it. This includes locating and restoring missing values, coping with missing column names, and ensuring proper data parsing and loading.

**Weeks 7-8:** Developing a data visualization project for a real-world dataset using Python and Excel.

SQL principles and data analysis applications, such as creating and querying databases, changing tables, and connecting tables.

Working on a real-world data analysis assignment with SQL.

**Resources:** DataCamp assignments: Joining data in SQL, Revise courses Intermediate SQL queries, Introduction to Joins, Outer joins and cross joins from datatcamp.

**Application:** Joining multiple tables using SQL based on different criteria, wrote various SQL queries to find different outputs of given dataset.

Using python and Pandas package to read preprocessed dataset: Setting column label for a given dataset, counting and identifying missing values, querying the dataset, etc.

**Weeks 9–10**: The fundamentals of machine learning, including as decision trees, clustering, and supervised and unsupervised learning, as well as data types, data preprocessing, and Python data processing.

**Resources:** Class lectures, Andreas Müller and Sarah Guido's "Introduction to Machine Learning with Python”, Datacamp assignment: Connecting data in tableau, Data story telling by Tableau, Data driven decision making in business.

**Application:** Creating scatterplot using Tableau, implementing k-means clustering using both Tableau and RapidMiner software and finding the difference between them.

Using RapidMainer creating a linear regression modeling of a training dataset, creating a prediction of a prediction dataset, creating a decision tree, and logistic regression modeling.

Making plot using single variable spatial data: Reading and cleaning selected dataset, creating donut plot, stacked bar, bar chart using colors for the viewers for visual comparison in Jupyter notebook.

**weeks 11 and 12**: Practice machine learning, utilize Python to work on a practical machine learning project.Review the time series forecasting supplemental materials.

**Resources:** Class lectures, books, online video tutorials.

**Application:** Making scatterplots of four variables of the selected dataset using dimension embedding in Jupyter notebook.

Making a line chart and a trending model of a selected dataset for analysis using Tableau, deciding tree model using RapidMiner to identify the key factors for customer churn from a given dataset, making regression model and prediction model for further analysis.

Creating a data visualization to find which factors boost the likelihood of car accidents in New York city and their potential effect on public transportation use by using data visualization using data selection, reading and cleaning of the data, dataset statistical description, and visualization.

**Log:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Week** | **Date** | **Topic** | **Materials** | **Application** | **Task completion status** |
| 1-2 | 1/17 to 1/29 | Overview of Data Analysis | Class lectures, online video tutorials, books |  | Completed |
| 3-4 | 2/1 to 2/12 | Analysis of Data with Python and Excel. | DataCamp assignment: Data analysis in excel, class lectures, review some materials from previous coursework TECH FOUNDATIONS CIS, code with mosh | Using Excel, examine a dataset of student responses. It goes over how to clean text data, use VLOOKUP, and get minimum, maximum, and average values. The analysis provides information on the dataset. | Completed |
| 5-6 | 2/15 to 3/5 | Basic statistical analysis concepts such as mean, median, mode, standard deviation, and hypothesis testing are covered. Data visualization with pandas, as well as pandas reading. | Class lectures, books, online video tutorials | Examine a selected dataset by cleaning and preparing the dataset, which includes handling missing values, column names, and data parsing verification by using excel and python | Completed |
| 7-8 | 3/8 to 3/26 | Developing a data visualization project in Python and Excel, understanding SQL principles and data analysis applications, and working on a real-world SQL data analysis assignment | DataCamp assignments: Joining data in SQL, Revise courses: Intermediate SQL queries, Introduction to Joins, Outer joins and cross joins from datatcamp. | Using SQL to join multiple tables and write queries to analyze a given dataset, as well as using Python and the Pandas package to read and preprocess the data. |  |
| 9-10 | 3/29 to 4/9 | Machine learning fundamentals (decision trees, clustering, supervised and unsupervised learning) and data processing fundamentals (data types, preprocessing, and Python processing). | DataCamp, class lectures, books | Using Tableau and RapidMiner, create scatterplots and implement k-means clustering, create multiple models, and plot single variable geographic data in Jupyter Notebook. | Completed |
| 11-12 | 4/12 to 4/16 | Practice machine learning, utilize Python to work on a practical machine learning project.Review the time series forecasting supplemental materials. | Class lectures, books, online video tutorials | Creating scatterplots in Jupyter Notebook, models in Tableau and RapidMiner to analyze datasets, and a data visualization to uncover factors influencing car accidents in New York City and public transportation use. | Completed |

**Reflections:** The goal of this design was “Developing your data analysis skills, including data collection, cleaning, analysis, and visualization”. After going through the rigorous schedule for three months, it can be said that the goal is accomplished. Also, there is a presence of Evidence of consistent practice and progress in the listed abilities (data collection, cleaning, analysis, and visualization).

Although every task is completed, there are some fluctuations. For example, there were some tasks regarding data visualization and analyzing data by using RapidMiner, those tasks were done very hastily because of the pressure of the deadlines. As a result, there is some lacking in the fluency of using these tools. To overcome this fact, there is a need of more frequent practice.

In completing these tasks, the biggest challenge was searching for the correct materials. To find the exact way to solve a problem usually takes the lion's share of the time. There were no other options than to devote a significant amount of time to it. Although it was tiresome labor, it helped to develop a habit of unwavering dedication to solving a problem.

Finally, it needs to state that sticking to this study plan helped to reach the goal and supported to become habituated to effective time management.