INFO8066 Data Analytics

Assignment 3

Due Date: November 19th, 2023 – 11:59 pm

Part 1: Long Answer Questions

a) Describe two common methods for handling missing data in Pandas DataFrames and provide an example for each.

b) Explain the significance and techniques of handling missing data in a real-world data analysis scenario. Discuss the challenges posed by missing data, the potential implications of different imputation methods, and the importance of assessing and documenting the handling of missing data in a data analysis project. Provide practical examples to illustrate the impact of various missing data strategies on the quality and reliability of results.

c) Explain the differences between Matplotlib, Seaborn, and Altair for data visualization in Python. Provide basic use cases and simple examples to help beginners understand when to use each of these libraries for creating visualizations.

Part 2: Data Cleaning

Download the Uncleaned_DS_Jobs.csv file. Use appropriate data cleaning techniques and methods in Pandas to convert the raw file into the clean data structure as show in the reference file "Cleaned_DS_Jobs.csv" file. Review each column and look at the data and if it makes sense.

Hints:

- Missing values are known as -1 in this dataset
- Delete columns that are not in the Cleaned version.
- In the salary_estimate column, delete the text (Glassdoor est.)
- In the company_name column, delete everything after the "\n"
- Review the location column and look for inconsistencies: You might need to split the column into 3 parts (State, location, City)
- In the industry column, replace -1 with n/a . Note this is only for the industry column. For other columns -1 will mean something else and hence can be dropped or filled in with other values
- Split the salary column into 3 parts, minimum, average, maximum salary
- From the job description column, extract the most popular skills (Python, excel, tableau, power_bi, sql,aws) and create separate columns for each with values 1/0 . 1 if the skill is

found in the corresponding job description and 0 if the skill is not found in the corresponding job description.

The provided hints address some of the specific cleaning tasks, but additional cleaning steps may be required. The end result should match the structure and cleanliness of the reference file.

Export/save file to a .csv using pandas.

Part 3: Exploratory data analysis and data visualization

Use any graphing/plotting libraries mentioned in the class to answer the following questions. Show all your work in the .ipynb file

a) Rank the top 10 best companies to work with according to ratings
b) Identify industries and Sectors with the highest demand for Data Science Professionals
c) Determine the average and maximum salary per sector
d) What are the top 5 US cities with most job listings
e) What are the most in demand industries in the job search and what is the average salary of each
f) Obtain the avg. salary based on the seniority of the roles
g) Which industries use AWS?

Part 4 : Conclusion

a) What story does this data tell us?
b) What is your conclusion based on your analysis? As a college student aspiring to work in the data field, what do your next steps look like after analyzing this dataset? *(Note this is an actual data scrap from Glassdoor hence the content of this data is accurate)*

Part 5 : Data Merging

Imagine you are working for a data analytics consulting firm, and you have been tasked with creating a dataset that complements the provided Glassdoor jobs postings dataset and perform further analysis.

a) Create a Custom Dataset: You can search for publicly available datasets online or create your own dummy dataset in excel using RANDOM values. Ensure that your dataset has a common 'key' value with "Cleaned_DS_Jobs" so that you can merge both datasets. Your custom dataset should have atleast 20 rows of data with atleast 3 columns.
Sample:

b) Merge the new dataset that you created with the “Cleaned_DS_Jobs” using the following techniques:
    a. Left
    b. Right
    c. Outer
    d. Inner

c) Answer the questions associated with each type of Join:
    a. Left
        i. What does this do when merging our two datasets? Provide explanation with example.
    b. Right
        i. Explain how this is different from Left join by analyzing the combined dataset
        ii. When is it appropriate to use a Right Join? Provide an example associated with the Glassdoor Dataset
    c. Outer Join
        i. When is this typically used in data merging?
        ii. Provide an example of a situation where you would choose an Outer Join over other types of joins.
    d. Inner Join
        i. Give an example of a scenario where an Inner Join is the most suitable merging technique.
        ii. What are the potential challenges or drawbacks of using an Inner Join?