Health Data Science: Homework 9
Karis Van Noord

1. How do you import the pandas package in Python?
    a. To import the pandas package in Python you input the following code: import pandas as pd.
2. What function do you use to read a CSV file in pandas?
    a. To read a CSV file in pandas, you input the following code: ‘pandas.read_csv’.
3. How do you display the first 5 rows of a DataFrame in pandas?
    a. To display the first five rows of data you input ‘head’.
4. How do you calculate the mean of a column named Age in a DataFrame named df?
    a. To calculate the mean of a column named Age in the dataframe, input the following code: df[‘Age’].mean()
5. How do you calculate the median of a column named Salary in a DataFrame named df?
    a. To calculate the median of a column named Salary in the dataframe, input the following code: df[‘Salary’].median()
6. How do you calculate the standard deviation of a column named Score in a DataFrame named df?
    a. To calculate the standard deviation of the column named score in the dataframe, input the following code: df[‘Score’].std()
7. How do you find the number of missing values in each column of a DataFrame named df?
    a. To find the number of missing values in each column of the dataframe, input the following code: df.isna().sum()
8. How do you calculate the correlation between two columns, Age and Salary, in a DataFrame named df?
    a. To calculate the correlation between two columns in the dataframe, input the following code: df[‘Age’].corr(df[‘Salary’])
9. How do you select a subset of a DataFrame df where the column Age is greater than 30?
    a. To select a subset of the dataframe where the age column is greater than 30, input the following code: df[df[‘Age’] > 30]
10. How do you calculate the range (maximum - minimum) of a column named Score in a DataFrame named df?
    a. To calculate the range of a column called “Score”, input the following code: df[‘Score’].max() - df[‘Score’].min()
11. How do you group a DataFrame df by a column named Department and calculate the mean of Salary within each group?
    a. To group the data by a column named “Department”, input the following code: df.groupby(‘Department’)[‘Salary’].mean()
12. How do you group a DataFrame df by two columns, Department and Job Title, and count the number of rows within each group?

    a. To group the data by two distinct columns and count the rows in each, input the following code: df.groupby([‘Department’, ‘Job Title’]).size()

13. How do you use the groupby method to find the maximum Age in each Department in a DataFrame df?
    a. Using the groupby method, find the maximum age in each department by inputting the following code: df.groupby(‘Department’)[‘Age’].max()
14. How do you create a cross-tabulation table that shows the frequency count of Department (rows) and Job Title (columns) in a DataFrame df?
    a. To create a cross-tabulation table that shows the frequency count of rows and columns, input the following code: pd.crosstab(df[‘Department’], df[‘Job Title’])
15. How do you create a cross-tabulation table that shows the mean Salary for each combination of Department (rows) and Job Title (columns) in a DataFrame df?
    a. To create a cross-tabulation table that shows the mean salary for each combination of department and job title, input the following code: pd.crosstab(df[‘Department’], df[‘Job Title’], values = df [‘Salary’], aggfunc= ‘mean’).