Homework Assignment 1 Cheat Sheet – Exploratory Data Analysis

Dr. Gail Gilboa-Freedman
Dr. Naveh Eskinazi, Mr. Asaf Lev

PYTHON CODING

LIBRARY: PANDAS

Purpose	Command
Import Library	*import pandas as pd*
Assign the CSV file into a variable	*datafile = 'filename.csv'*
Read CSV file into a dataset	*dataset = pd.read_csv(datafile)*
Assign the dataset into a dataframe	*dataframe = pd.DataFrame(dataset)*
Show the dataframe content	*dataframe*
Number of unique values in a column	*unique = len(pd.unique(dataframe['column_name'])) / unique=dataframe['column_name'].nunique()*
Show the data of range of rows	*dataframe.iloc[index1:index2]*
Create a list of cases that meet a criterion (result: true / false)	*index = dataframe['column_name'] == criteria*
Show the actual data of the cases	*dataframe[index]*
Show the descriptive statistics of all numerical columns within the dataframe	*dataframe.describe()*
Show the number of cases for each unique value in a column	*dataframe['column_name'].value_counts()*

Groupby - Group a dataframe by “x” and get statistic “metric” on “y” *Pay attention: X,Y and Metric should be replaced with the actual value	*dataframe.groupby(“x”)[“y”].metric()* *For example:* *find the mean of milk products sold by region in wholesale_data i.e., the data used in the tutorial):* *dataframe.groupby(“Region”)[“Milk”].mean()*
Create a bar chart based on a grouped dataframe (Note: Bar plot can be created also on a non-grouped data frame as shown later)	*grouped_dataframe.plot(kind='bar')* *For example:* *Create a bar chart that will compare the mean of milk products sold in each region in wholesale_data:* *grouped_dataframe = dataframe.groupby(“Region”)[“Milk”].mean()* *grouped_dataframe.plot(kind='bar')* *Optional: add titles to the axes and to the chart*

LIBRARY: MATPLOTLIB

Purpose	Command
Import Library	*Import matplotlib as plt*
Import pyplot module	*Import matplotlib.pyplot as plt*
Visualize data of a column using Box & Whisker	*dataframe.boxpot(column=’column_name’)*
Create a histogram of a given column with bins defined as column’s number of unique values	*plt.hist(dataframe[‘column_name’], bins = unique)*
Define x-axis data	*x = dataframe [‘column_name’]*
Define y-axis data	*y = dataframe [‘column_name’]*

Create a bar chart	*plt.bar(x, y)*
Define x-axis title	*plt.xlabel('enter your text')*
Define y-axis title	*plt.ylabel('enter your text')*
Define chart title	*plt.title('enter your text')*

Good Luck!