DI Martin Uray
martin.uray@fh-salzburg.ac.at

Applied AI with Python
Assignment Sheet 2: Visualizing Categorical and Geographical Data

Version:
2024-11-12

To prepare for the next session of the course, students are expected to complete the following exercises. At the beginning of the session, a handout will be distributed containing a list of exercises. Students will be required to tick off their individual achievements as evidence of their work. During the session, students will be randomly selected from those present and will be asked to present (defend) their solutions. A brief interview will follow, during which the student's solution will be discussed. If any doubts arise regarding a student's presentation that they are unable to explain, no credit will be awarded for that exercise in the final grading. To ensure a collective learning experience, the instructor will also review and discuss the presented work with the plenum.

1 Categorical Data

In this part, you'll take a look at a publicly available data set containing data on the Titanic and its passengers - the Titanic data set. You are going to visualize the data describing the people aboard the vessel and find out more about their fates.

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Table 1: Meta Information for Titanic data set

Data Loading and Selection Methods

Ex. 9 **Data Loading and Selection Methods:** To complete the first assignment, you are asked to apply your knowledge in Python and pandas to the Titanic data set. Table 1 contains some meta information on the data set, providing some additional context for you. After you got to know more about the Titanic data set, answer the following questions using pandas data selection methods:

- How many passengers have embarked in Southampton?
- What was the price of the most expensive ticket? Find out the name(s) of the passenger(s) who bought them.
- Who were the youngest passengers? Find out their name(s) and age(s).
- Who was/were the oldest passenger(s) who died during the trip? Find out their name(s) and age(s).

Visualizing Categorical Data

Familiarize yourself with matplotlib and plotly to get to know some of the graphing options these libraries provide. The three exercises below are to be answered using plots. Don't forget to set appropriate values for title and axis labels and consider customizing the legend where applicable.

Ex. 10 Create a plot that gives us an idea how likely it was for passengers of different classes (1st, 2nd, 3rd) to survive.

Ex. 11 Visualizing passengers' journeys using their embarking locations and their survival status (think voter-flow analysis).

Ex. 12 Make up one additional question yourself and come up with a fitting plot to answer it!

2 Categorical Data

GeoPandas

Now, you will work with open source geographical data from naturalearthdata.com alongside GeoPandas, an open source project used to make working with geospatial data in python easier. It **extends the datatypes used by pandas** to allow spatial operations on geometric types.

Data Structures

Similarly to pandas, GeoPandas provides two data structures for working with geographic data:

- `GeoSeries`: A vector where each entry in the vector is a set of shapes corresponding to one observation. An entry may consist of only one shape (like a single polygon) or multiple shapes that are meant to be thought of as one observation (like the many polygons that make up the State of Hawaii or a country like Indonesia). Typical entries include Points or Multi-Points, Lines or Multi-Lines and Polygons or Multi-Polygons
- `GeoDataFrame`: A tabular data structure that contains one `GeoSeries` column that holds geospatial information. This `GeoSeries` is referred to as the "Geometry" of a `GeoDataFrame`. When a spatial method is applied to a `GeoDataFrame` (or a spatial attribute like area is called, these commands will act on the "Geometry" column.

Ex. 13 **Plot a Choropleth map focusing on a single continent:** Apply what you learned about how to create your own Choropleth maps by selecting a continent (not Antarctica, please ;-)) and plotting the percentage of the population using the Internet for each country displayed. Since this dataset includes historical data, you need to select a specific year of your interest as well before being able to plot the data on a map!

Ex. 14 **Plot the Internet usage of a single country over time:** Next make use of the historical data provided in the dataset linked above and single out a specific country. Visualize how the Internet usage over time changed for its population!

Ex. 15 **Create one additional Choropleth Map of your choosing:** Find some other information to overlay on a map of a country, continent or other part of the world. The easiest way to achieve this is by finding open source data that is geocoded in some way (e.g. with ISO country codes) and joining it onto the dataframes available from naturalearthdata.com containing geographic data. One possible source of such data is data.worldbank.org.