DSO 545: Statistical Computing and Data Visualization

Abbass Al Sharif

Fall 2019

Lab 7: Data Wrangling

1. Data Science Tools: `map()`, `apply()`, and `lambda` functions

1. Load the `titanic.csv` dataset into a pandas DataFrame.
2. Use the map() function to create a new variable call "Sex_Numeric" (0 for males, and 1 for females).
3. What is the percentage of females in the Titanic dataset?
4. Create a new variable called "Fare_ceil" to show the fare value of the trip rounded up.
5. Load the `drinks.csv` dataset into a dataset called `drinks`.
6. Find the max beer, spirit, and wine servings among all countries.
7. Find the max serving among beer, spirit, and wine for each country.
8. Find the max category of serving among beer, spirit, and wine for each country.
9. Creata a function that take a input value x and returns it four-fold.
10. Create a new column in the drinks dataset `beer_4fold` which multiplies the value of the beer servings in each country by 4.
11. Use a lambda function to answer the previous question, i.e. don't use the function `four_fold()`.

2. Data Wrangling

Filtering Data

12. Find all passangers who are above 30 years old.
13. Find all female passangers who are above 30 years old.

Selecting Variables

14. Create a dataframe which has only two columns: `PassengerId`, `Survived`, and `Cabin`.
15. Create a dataframe that has all variables in the dataset except the "Cabin" variable.

Arranging Data

16. Find all female passangers who are above 30 years old. The resulting dataframe should have three columns: PassengerId, Survived, Age. The dataframe should be arranged by age in descending order.

Grouping and Summarizing Data

17. Find the average age of both male and female passangers.
18. Find the median fare for passengers according to their class.
19. Find the average and median, and the difference between mean and median age of both male and female passangers.
20. Find the average age for males and females who survived the Titanic disaster.
21. Find the median fare for passengers embarked from different ports and among different classes.

Mutating Data

22. Create a new column (age_cat) in the dataset to based on the age variable ("young" if age <=50 otherwise "older").
23. Create a new column (age_cat1) in the dataset to based on the age variable ("young" if age <=20, "mature" if 20<age<=50 otherwise "older").
24. Create the following Tree map using the `mpg.csv` dataset and `squarify` Python package.