Lab 7

The Cars93 dataset contains data from 93 cars that were on sale in the USA in 1993. The dataset has 93 rows listing various car models along with their features and specifications. The columns of this dataset---there are 27 of them---comprise various attributes, e.g., Manufacturer, Type, Price, mileage, horsepower etc. Refer to the ***Appendix*** *for more information.*

```
# Getting started with pandas:
https://pandas.pydata.org/docs/getting_started/index.html
# API reference: https://pandas.pydata.org/docs/reference/index.html
# Reading the data from a csv file and storing it in a panda dataframe
import panda as pd
titanic = pd.read_csv("data/titanic.csv")
```

Sorting, grouping and printing

1. List all the manufacturers mentioned in the Cars93 data set in alphabetical order.
   [Hint: `df.sort_values(['col1', 'col2'])`
   `df.sort_values('col1', ascending=False)`]
2. Print all details of the costliest car of each of the 'Types'.
   [Hint: `df[df[‘col1’] == df[‘col1’].max()]`]

3. Write a function that asks you to enter the name of a manufacturer and print all the models produced by that manufacturer. Print an appropriate error message if the input is invalid.
   [Hint: `df[df[‘col1’]== ‘string1’]`]
4. Print the total count of cars per manufacturer. Print it in the following format:

   ```
   -------------------------------------
    Manufacturer      Count
   --------------------------------------
           Mfr1           N1
           Mfr2           N2
           Mfr3           N3
   ```

5. List all the non-USA "Small" ('Type') cars ('Make') in alphabetical order.

6. Group the cars by ‘Manufacturer’ and find the average price for each of the manufacturers by averaging the ‘Price’ of all the models produced by a particular manufacturer and then print all the ‘Manufacturers’ sorted (in ascending order) by average price.
7. Find the car models that are “Midsize” (‘Type’) and cost less than the average cost of all the cars (of all the types).
8. Find the car models that have MPG.city and MPG.highway higher than the average of MPG.city and average of MPG.highway, respectively. Here, average should be calculated over all the cars.

[Hint: `df_subset = df[(df[`col1’] >= val1) & (df[‘col2’] >= val2)]`]

Data derivation, computation and comparison

```
# creating new columns from existing ones
air_quality["ratio_paris_antwerp"] = (air_quality["station_paris"] /
air_quality["station_antwerp"])
```

9. Calculate the ratio of Max.Price and Min.Price for each of the car models and add a new column called ‘Max_min_ratio’ to the imported dataframe.
   a. Print the cars for whom the ratio is equal to one.
   b. Find the manufacturer whom the gap between largest and smallest max_min_ratio value is highest. Also, print the corresponding ratio values.
   c. Print the car models (manufacturer and model) for whom the max_min_ratio is within ±1% of the mean of Max_min_ratio.
10. Calculate the average MPG ((MPG.city + MPG.highway)/2) for each of the car models and add a new column called ‘AverageMPG’ to the updated dataframe
    a. List the top 5 cars based on ‘AverageMPG’.
    b. Sort the “Midsize” cars by ‘AverageMPG’.
    c. Print all the non-USA cars (manufacturer and model) in alphabetical order whose ‘AverageMPG’ is higher than the average of the ‘AverageMPG’ column.
11. Create a new identifier for the cars listed in the dataset by concatenating the first three alphabets of ‘Manufacturer’ with the first three alphabets of ‘Model’ followed by the horsepower. Use hyphens to indicate the concatenation. For example, the identifier for the first row is “Acu-Int-140”.
    a. Add a new column called ‘Identifier’ to the updated dataframe and generate identifiers for each of the car models.
    b. Save the updated dataframe as a new csv file by suffixing your roll number to the filename. For example, if your roll number is b20500, then the file name should be “Cars93b20500.csv”. Also, note that the updated dataframe has now three new columns, namely, ‘Max_min_ratio’, ‘AverageMPG’, and ‘Identifier’.

To verify that load both the csv files and list the columns that are not present in the original csv file, i.e., “Cars93.csv”.

12. Compare the average prices (average price is calculated after grouping them based on the ‘Type’) of different types (“Small”, “Midsize” etc.) of non-USA cars with that of the USA. Which one is cheaper in each category?

Statistical analysis

```
# finding correlation between col1 and col2
df[‘col1’].corr(df[‘col2’])
```

13. Find the correlation of the area (‘Length’ x ‘Width’) with (a) ‘Passengers’, (b) ‘Luggage.room’, (c) ‘Rear.seat.room’, and (d) ‘Weight’. For each of the cases, did the sign of the *correlation coefficient* match your expectations?
14. Find the correlation between (a) ‘EngineSize’ and ‘Horsepower’, (b) ‘Cylinders’ and ‘EngineSize’, (c) ‘Horsepower’ and ‘RPM’, (d) ‘EngineSize’ and ‘Fuel.tank.capacity’. For each of the cases, did the sign of the correlation coefficient match your expectations?
15. Plot histograms of the following columns with a bin size(or width) mentioned next to the column name after converting the values into the unit mentioned inside the brackets
    [Hint: `df[‘col1’].hist(bins = <num_of_bins>)` *`# you need to find the number of bins from the bin size mentioned below`*]
    a. Price, 5 (unit is Lakhs) [Assume $1 = INR 73.5]
    b. AverageMPG, 2.5 (unit is Kilometres per liter)
    c. Horsepower, 20 (unit is Kilowatt)
    d. Area, 1 (unit is sq. metres)
    e. Weight, 250 (unit is kg)

Appendix

Source: https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/Cars93.html (retrieved on 1 Apr 2021)

Details

Cars were selected at random from among 1993 passenger car models that were listed in both the Consumer Reports issue and the PACE Buying Guide. Pickup trucks and Sport/Utility vehicles were eliminated due to incomplete information in the Consumer Reports source. Duplicate models (e.g., Dodge Shadow and Plymouth Sundance) were listed at most once.

Further description can be found in Lock (1993).

Format

This data frame contains the following columns:

Manufacturer: Manufacturer

Model: Model.

Type: a factor with levels "Small", "Sporty", "Compact", "Midsize", "Large" and "Van".

Min.Price: Minimum Price (in \$1,000): price for a basic version.

Price: Midrange Price (in \$1,000): average of Min.Price and

Max.Price: Maximum Price (in \$1,000): price for “a premium version”.

MPG.city: City MPG (miles per US gallon by EPA rating).

MPG.highway: Highway MPG.

AirBags: Air Bags standard. Factor: none, driver only, or driver & passenger.

DriveTrain: Drive train type: rear wheel, front wheel or 4WD; (factor).

Cylinders: Number of cylinders (missing for Mazda RX-7, which has a rotary engine).

EngineSize: Engine size (litres).

Horsepower: Horsepower (maximum).

RPM: revs per minute at maximum horsepower

Rev.per.mile: Engine revolutions per mile (in highest gear).

Man.trans.avail: Is a manual transmission version available? (yes or no, Factor).

Fuel.tank.capacity: Fuel tank capacity (US gallons).

Passengers: Passenger capacity (persons)

Length: Length (inches).

Wheelbase: Wheelbase (inches).

Width: Width (inches).

Turn.circle: U-turn space (feet).

Rear.seat.room: Rear seat room (inches) (missing for 2-seater vehicles).

Luggage.room: Luggage capacity (cubic feet) (missing for vans).

Weight: Weight (pounds).

Origin: Of non-USA or USA company origins? (factor).

Make: Combination of Manufacturer and Model (character).

Source

Lock, R. H. (1993) 1993 New Car Data. Journal of Statistics Education 1(1). doi: 10.1080/10691898.1993.11910459

References

Venables, W. N. and Ripley, B. D. (1999) Modern Applied Statistics with S-PLUS. Third Edition. Springer.