Aditya Parey

ALY 6010 Summer 2024

R Practice 1

Practice Module 1

Overview:

The 'LungCapDataCSV.csv' dataset, which contains variables pertaining to lung capacity, smoking status, and demographic data, is the subject of the analysis. The variables of interest, the data cleansing procedure, the preliminary analysis processes, the outcomes, and the visualisations produced are all documented in the report.

Variables:

This report's main variables under analysis are:  
  
Lung Capacity (lung\_cap)

Age (age)  
Gender  
Smoking Status

Data Cleaning:

Using clean\_names from the janitor package to standardize the variable names.

A close-up of words

Description automatically generated

Data types:  
In order to express categorical data, the gender variable was transformed into a factor.  
To better represent its numerical nature, the age field was transformed to an integer.

A screenshot of a computer code

Description automatically generated

Initial Analysis with frequency tables and cross-tabulation:

A screenshot of a computer code

Description automatically generated

A screenshot of a data sheet

Description automatically generated

Analysis with histograms and comparisons across variables:

A graph of a frequency

Description automatically generated

A green and black graph

Description automatically generated

A graph of a person and person

Description automatically generated

A graph of a number of people

Description automatically generated with medium confidence

Conclusion:

This paper provides a thorough examination of the 'LungCapDataCSV.csv' dataset with an emphasis on age, gender, smoking status, and lung capacity. After being cleaned, the data was subjected to frequency tables, cross-tabulations, and histograms and bar graphs for visualisation. The knowledge gathered from this analysis may help guide future investigations into the variables influencing smoking behaviour and lung capacity.