Tutorial 7:

Introduction to Data Science

Exploratory Data Analysis

Ernst C. Wit

November 4, 2024

Exercises

1. **Categorical variable.** The housing market in Melbourne has been quite active. Load the data `melbournehousing.csv` from iCorsi containing more than 13,500 property sales. In this question we are interested in `Type` of property sold:

   - br - bedroom(s);
   - h - house,cottage,villa, semi,terrace;
   - u - unit, duplex;
   - t - townhouse;
   - dev site - development site;
   - o res - other residential.

   (a) Evaluate the number of the various types of house sales.

   (b) Evaluate the relative frequency of the various type of house sales.

   (c) What is the most popular type of house?

2. **Continuous variable.** In the R-package `HistData` you can find the `Arbuthnot` dataset that contains information about births in London between 1629 and 1710. The data actually describe the number of christenings, meaning that still births and peri-natal deaths were discarded. John Arbuthnot collected these data in order to evaluate whether the birth ratio for boys and girls was really 1 or not.

   In this exercise we focus on the variable `Ratio` that describes the ratio of male over female births. The data can also be found on iCorsi if you want to make use of Python.

   (a) Use the various summary statistics for location in order to identify the typical value for the ratio of female over male births.

   (b) What is the variation of the birth ratio?

   (c) Use various visual displays to describe the spread and skew of the birth ratio.

   (d) Given your various summaries try to answer Arbuthnot's original question.

3. **Relationship between 2 categorical variables.** We return to the Melbourne housing dataset. In this question, we want to evaluate the relationship between the type of house sold and the Region.

   (a) Use cross-tabulation to explore the relationship between the region and the type of house sold.

(b) Normalize the table in an appropriate way in order to see the relative percentages of each house type in each region.

(c) In which region is the townhouse most popular?

4. **Relationship categorical and continuous variable.** In the history of data visualization, Florence Nightingale is best remembered for her role as a social activist and her view that statistical data, presented in charts and diagrams, could be used as powerful arguments for medical reform. After witnessing deplorable sanitary conditions in the Crimea, she wrote several influential texts (Nightingale, 1858, 1859), including polar-area graphs (sometimes called Coxcombs or rose diagrams), showing the number of deaths in the Crimean from battle compared to disease or preventable causes that could be reduced by better battlefield nursing care.

   Load the data as the dataframe `Nightingale` from the R-package `HistData` or directly from iCorsi.

   (a) Make a data visualization of the relationship between the number of deaths due to preventable or mitagable diseases and the month of the year.

   (b) Make a data visualization of the relationship between the number of deaths due to battle and the month of the year.

   (c) What can you conclude from these two plots?

5. **Multiple continuous variables.** The Spanish Armada (Spanish: Grande y Felicisima Armada, literally "Great and Most Fortunate Navy") was a Spanish fleet of 130 ships that sailed from La Coruna in August 1588. During its preparation, several accounts of its formidable strength were circulated to reassure allied powers of Spain or to intimidate its enemies. One such account was given by Paz Salas et Alvarez (1588) and available in the R-package `HistData` as the dataframe `Armada`. The data is also available on iCorsi. In this question, we will explore the relationship between various continuous variables.

   By the way, although the intent was bring the forces of Spain to invade England, overthrow Queen Elizabeth I, and re-establish Spanish control of the Netherlands, the Armada was not as fortunate as hoped: it was all destroyed in one week's fighting by Francis Drake.

   (a) What is the relationship between the number of ships from the various regions and the amount of artillery it was carrying?

   (b) What is the relationship between the amount of artillery and the amount of cannon balls? How many cannon ball does each piece of artillery have available?

   (c) In R, use the function `pairs` to plot the pairwise relationship between all the variables simultaneously.