Lec 5
Practical case studies: Descriptive statistics

Aug 01

Content

- Practical case studies: Descriptive statistics
  - handling missing data
  - Anomalies
  - feature selection
- Introduction to scikit-learn
  - Estimators
  - Decision Trees

What is Data Science Pipeline

- Data driven decisions.
- Analyse raw data to produce actionable results.
- Skills required for DS:
  - Hacking (Coding)
  - Maths & Stats
  - Domain Knowledge
- Data mining: Extract knowledge from data.
  - Correlations
  - ML Models
  - From Linear Regression to complex NN.

NHANES case study

- importing the libraries in python.
  - %matplotlib inline
  - import matplotlib.pyplot as plt
  - import seaborn as sns
  - import pandas as pd
  - import numpy as np

- 1. Load the dataset
  - df = pd.read_csv("nhanes_2015_2016.csv")

1.1 Frequency tables

- determine the number of times that each distinct value of a variable occurs in a data set.
  - df.DMDEDUC2.value_counts()

```
4.0    1621
5.0    1366
3.0    1186
1.0     655
2.0     643
9.0       3
Name: DMDEDUC2, dtype: int64
```

1.2 Missing values

- ‘value_counts’ method excludes missing values.
  - print(df.DMDEDUC2.value_counts().sum())
  - print(da.shape)
- Another way is to locate the null values
  - pd.isnull(d.DMDEDUC2).sum()

```
5474
5474
(5735, 28)
```

1.3 Re-label

- replace integer codes with a text label
  - df["DMDEDUC2x"] = df.DMDEDUC2.replace({1: "<9", 2: "9-11", 3: "HS/GED", 4: "Some college/AA", 5: "College", 7: "Refused", 9: "Don't know"})
  - df.DMDEDUC2x.value_counts()

```
Some college/AA    1621
College            1366
HS/GED             1186
<9                  655
9-11                643
Don't know            3
Name: DMDEDUC2x, dtype: int64
```

1.4 Creating a category

- Missing category
  - df["DMDEDUC2x"] = df.DMDEDUC2x.fillna("Missing")
  - x = df.DMDEDUC2x.value_counts()
  - x / x.sum()

```
Some college/AA    0.282650
College            0.238187
HS/GED             0.206800
<9                 0.114211
9-11               0.112119
Missing            0.045510
Don't know         0.000523
Name: DMDEDUC2x, dtype: float64
```

1.5 Numerical Summaries

- describe() method
  - df.BMXWT.dropna().describe()

```
count    5666.000000
mean       81.342676
std        21.764409
min        32.400000
25%        65.900000
50%        78.200000
75%        92.700000
max       198.900000
Name: BMXWT, dtype: float64
```