4_descriptive_statistics.ipynb

File Edit View Insert Runtime Tools Help Cannot save changes

Share

+ Code + Text Copy to Drive

RAM Disk

Editing

```
[15] import pandas as pd
```

Import data

```
[19] # Import data from GitHub (or from your local computer)
     df = pd.read_csv("https://raw.githubusercontent.com/kirenz/datasets/master/wage.csv")
     df
```

	Unnamed: 0	year	age	maritl	race	education	region	jobclass	health	health_ins	logwage	wage
0	231655	2006	18	1. Never Married	1. White	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.318063	75.043154
1	86582	2004	24	1. Never Married	1. White	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	2. No	4.255273	70.476020
2	161300	2003	45	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.875061	130.982177
3	155159	2003	43	2. Married	3. Asian	4. College Grad	2. Middle Atlantic	2. Information	2. >=Very Good	1. Yes	5.041393	154.685293
4	11443	2005	50	4. Divorced	1. White	2. HS Grad	2. Middle Atlantic	2. Information	1. <=Good	1. Yes	4.318063	75.043154
...	...	...	...	...	...	...	...	...	...	...	...	...
2995	376816	2008	44	2. Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	5.041393	154.685293
2996	302281	2007	30	2. Married	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	2. >=Very Good	2. No	4.602060	99.689464
2997	10033	2005	27	2. Married	2. Black	1. < HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	2. No	4.193125	66.229408
2998	14375	2005	27	1. Never Married	1. White	3. Some College	2. Middle Atlantic	1. Industrial	2. >=Very Good	1. Yes	4.477121	87.981033
2999	453557	2009	55	5. Separated	1. White	2. HS Grad	2. Middle Atlantic	1. Industrial	1. <=Good	1. Yes	4.505150	90.481913

3000 rows × 12 columns

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3000 entries, 0 to 2999
Data columns (total 12 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Unnamed: 0  3000 non-null   int64
 1   year        3000 non-null   int64
 2   age         3000 non-null   int64
 3   maritl      3000 non-null   object
 4   race        3000 non-null   object
 5   education   3000 non-null   object
 6   region      3000 non-null   object
 7   jobclass    3000 non-null   object
 8   health      3000 non-null   object
 9   health_ins  3000 non-null   object
 10  logwage     3000 non-null   float64
 11  wage        3000 non-null   float64
dtypes: float64(2), int64(3), object(7)
memory usage: 281.4+ KB
```

Measures of central tendency

First of all we obtain some common statistics per variable.

```
[20] # mode
     df['age'].mode()
```

```
0    40
dtype: int64
```

```
[21] # calculation of the mean (e.g. for age)
     df["age"].mean()
```

```
42.41466666666667
```

```
[22] # calculation of the mean (e.g. for age) and round the result
     round(df["age"].mean(), 2)
```

```
42.41
```

```
[23] # calculation of the median (e.g. for age)
     df["age"].median()
```

```
42.0
```

Measures of dispersion

```
[25] # quantiles
     df['age'].quantile([.25, .5, .75])
```

```
0.25    33.75
0.50    42.00
0.75    51.00
Name: age, dtype: float64
```

```
[26] # Range
df['age'].max() - df['age'].min()
```

```
62
```

```
[27] # standard deviation
round(df['age'].std(),2)
```

```
11.54
```

```
[31] # variance
df['age'].var()
```

```
133.22712726464468
```

Summary statistics

```
[28] # summary statistics for all numerical columns
round(df.describe(),2)
```

	Unnamed: 0	year	age	logwage	wage
count	3000.00	3000.00	3000.00	3000.00	3000.00
mean	218883.37	2005.79	42.41	4.65	111.70
std	145654.07	2.03	11.54	0.35	41.73
min	7373.00	2003.00	18.00	3.00	20.09
25%	85622.25	2004.00	33.75	4.45	85.38
50%	228799.50	2006.00	42.00	4.65	104.92
75%	374759.50	2008.00	51.00	4.86	128.68
max	453870.00	2009.00	80.00	5.76	318.34

Compare summary statistics for specific groups in the data:

```
[29] # summary statistics by groups
df['age'].groupby(df['education']).describe()
```

education	count	mean	std	min	25%	50%	75%	max
1. < HS Grad	268.0	41.794776	12.611111	18.0	33.0	41.5	50.25	75.0
2. HS Grad	971.0	42.217302	12.023480	18.0	33.0	42.0	50.00	80.0
3. Some College	650.0	40.887692	11.523327	18.0	32.0	40.0	49.00	80.0
4. College Grad	685.0	42.773723	10.902406	22.0	34.0	43.0	51.00	76.0
5. Advanced Degree	426.0	45.007042	10.263468	25.0	38.0	44.0	53.00	76.0