Ben Deatsman

Module 3 Lab

Subsetting Data

In [4]:
```
import pandas as pd
df = pd.read_csv('heart.csv')
```

In [10]:
```
df[df['max_hr'] <= 100]
```

Out[10]:

	age	sex	chest_pain	rest_bp	chol	fbs	rest_ecg	max_hr	exang	old_peak	slope	ca	thal	disease
72	62	Male	asymptomatic	120	267	0	normal	99	1	1.8	2	2.0	reversable	1
114	62	Female	nonanginal	130	263	0	normal	97	0	1.2	2	1.0	reversable	1
154	64	Male	asymptomatic	120	246	0	left ventricular hypertrophy	96	1	2.2	3	1.0	normal	1
175	57	Male	asymptomatic	152	274	0	normal	88	1	1.2	2	1.0	reversable	1
223	53	Male	asymptomatic	123	282	0	normal	95	1	2.0	2	2.0	reversable	1
244	60	Female	nonanginal	120	178	1	normal	96	0	0.0	1	0.0	normal	0
245	67	Male	asymptomatic	120	237	0	normal	71	0	1.0	2	0.0	normal	1
296	59	Male	asymptomatic	164	176	1	left ventricular hypertrophy	90	0	1.0	2	2.0	fixed	1

In [14]:
```
df[df['max_hr'] <= 100].shape
```

Out[14]:
```
(8, 14)
```

There are 8 observations

In [25]:
```
df[(df['rest_bp'] >= 120) & (df['sex'] == 'Female')]
```

Out[25]:

	age	sex	chest_pain	rest_bp	chol	fbs	rest_ecg	max_hr	exang	old_peak	slope	ca	thal	disease
4	41	Female	nontypical	130	204	0	left ventricular hypertrophy	172	0	1.4	1	0.0	normal	0
6	62	Female	asymptomatic	140	268	0	left ventricular hypertrophy	160	0	3.6	3	2.0	normal	1
7	57	Female	asymptomatic	120	354	0	normal	163	1	0.6	1	0.0	normal	0
11	56	Female	nontypical	140	294	0	left ventricular hypertrophy	153	0	1.3	2	0.0	normal	0
18	48	Female	nonanginal	130	275	0	normal	139	0	0.2	1	0.0	normal	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
286	58	Female	asymptomatic	170	225	1	left ventricular hypertrophy	146	1	2.8	2	2.0	fixed	1
291	55	Female	nontypical	132	342	0	normal	166	0	1.2	1	0.0	normal	0
294	63	Female	asymptomatic	124	197	0	normal	136	1	0.0	2	0.0	normal	1
297	57	Female	asymptomatic	140	241	0	normal	123	1	0.2	2	0.0	reversable	1
301	57	Female	nontypical	130	236	0	left ventricular hypertrophy	174	0	0.0	2	1.0	normal	1

78 rows × 14 columns

In [27]:
```
df[(df['rest_bp'] >= 120) & (df['sex'] == 'Female')].shape
```

Out[27]:
```
(78, 14)
```

There are 78 observations

In [30]:
```
df[(df['fbs'] == 1) & (df['chest_pain'] == 'nontypical') & (df['disease'] == 1)]
```

Out[30]:

	age	sex	chest_pain	rest_bp	chol	fbs	rest_ecg	max_hr	exang	old_peak	slope	ca	thal	disease
261	58	Female	nontypical	136	319	1	left ventricular hypertrophy	152	0	0.0	1	2.0	normal	1

In [32]:
```
df[(df['fbs'] == 1) & (df['chest_pain'] == 'nontypical') & (df['disease'] == 1)].shape
```

Out[32]:
```
(1, 14)
```

There is 1 row and 14 columns

Manipulating Data

In [38]:
```
df.isnull().sum()
```

Out[38]:
```
age          0
sex          0
chest_pain   0
rest_bp      0
chol         0
fbs          0
rest_ecg     0
max_hr       0
exang        0
old_peak     0
slope        0
ca           4
thal         2
disease      0
dtype: int64
```

There are 4 missing values in the 'ca' column and 2 in the 'thal' column

In [49]:
```
df['ca'].fillna(df['ca'].mean(), inplace=True)
df['thal'].fillna(df['thal'].mode()[0], inplace=True)
```

In [51]:
```
df.isnull().sum()
```

Out[51]:
```
age          0
sex          0
chest_pain   0
rest_bp      0
chol         0
fbs          0
rest_ecg     0
max_hr       0
exang        0
old_peak     0
slope        0
ca           0
thal         0
disease      0
dtype: int64
```

In [57]:
```
risk = df['age'] / (df['rest_bp'] + df['chol'] + df['max_hr'])
df['risk'] = risk
df['risk'].max()
```

Out[57]:
```
0.18393782383419688
```

In [59]:
```
df['risk'].min()
```

Out[59]:
```
0.054104477611940295
```

In [62]:
```
vmap = {'left ventricular hypertrophy ': 'lvh', 'ST-T wave abnormality': 'stt_wav_abn'}
df['rest_ecg'] = df['rest_ecg'].replace(vmap)
```

In [80]:
```
df[df['rest_ecg'] == 'lvh']
```

Out[80]:

	age	sex	chest_pain	rest_bp	chol	fbs	rest_ecg	max_hr	exang	old_peak	slope	ca	thal	disease	risk
0	63	Male	typical	145	233	1	lvh	150	0	2.3	3	0.0	fixed	0	0.119318
1	67	Male	asymptomatic	160	286	0	lvh	108	1	1.5	2	3.0	normal	1	0.120939
2	67	Male	asymptomatic	120	229	0	lvh	129	1	2.6	2	2.0	reversable	1	0.140167
4	41	Female	nontypical	130	204	0	lvh	172	0	1.4	1	0.0	normal	0	0.081028
6	62	Female	asymptomatic	140	268	0	lvh	160	0	3.6	3	2.0	normal	1	0.109155
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
288	56	Male	nontypical	130	221	0	lvh	163	0	0.0	1	0.0	reversable	0	0.108949
290	67	Male	nonanginal	152	212	0	lvh	150	0	0.8	2	0.0	reversable	1	0.130350
293	63	Male	asymptomatic	140	187	0	lvh	144	1	4.0	1	2.0	reversable	1	0.133758
296	59	Male	asymptomatic	164	176	1	lvh	90	0	1.0	2	2.0	fixed	1	0.137209
301	57	Female	nontypical	130	236	0	lvh	174	0	0.0	2	1.0	normal	1	0.105556

148 rows × 15 columns

148 observations replaced 'left ventricular hypertrophy' with 'lvh'

Summarizing Data

In [101…
```
df['rest_bp'].groupby(df['sex']).mean()
df.groupby('sex', as_index=False).agg({'rest_bp': 'mean'})
```

Out[101…

	sex	rest_bp
0	Female	133.340206
1	Male	130.912621

In [99]:
```
df.groupby('sex', as_index=False).agg({'chol': ['mean', 'median']})
```

Out[99]:

	sex	chol	
		mean	median
0	Female	261.752577	254.0
1	Male	239.601942	235.0

In [111…
```
dfmale = df[df['sex'] == 'Male']
dfmale.groupby('age', as_index=False).agg({'chol': 'median'}).max()
```

Out[111…
```
age      77.0
chol    304.0
dtype: float64
```

77 year olds have the highest median cholestorol level.

In [120…
```
dfriskmean = df.groupby(['sex', 'age'], as_index=False).agg({'risk': 'mean'})
dfriskmean = dfriskmean.rename(columns={'risk': 'mean risk'})
dfriskmean
```

Out[120…

	sex	age	mean risk
0	Female	34	0.065385
1	Female	35	0.069583
2	Female	37	0.073267
3	Female	39	0.079549
4	Female	41	0.077815
...	...	...	...
68	Male	67	0.133246
69	Male	68	0.126461
70	Male	69	0.129603
71	Male	70	0.135125
72	Male	77	0.130288

73 rows × 3 columns

In [189…
```
dfriskmean.sort_values('mean risk', axis=0, ascending=False, kind='mergesoft').head(1)
```

Out[189…

	sex	age	mean risk
34	Female	76	0.16777

76 year old women have the highest risk value.

In [ ]:
```

```