project_24-3-2025

March 24, 2025

[1]:

```
#import the libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

[2]:

```
# loading the file
df =pd.read_csv('insurance.csv')
```

[3]:

```
df.shape
```

[3]:

```
(1338, 7)
```

[4]:

```
df.head()
```

[4]:

```
   age     sex     bmi  children smoker     region      charges
0   19  female  27.900         0    yes  southwest  16884.92400
1   18    male  33.770         1     no  southeast   1725.55230
2   28    male  33.000         3     no  southeast   4449.46200
3   33    male  22.705         0     no  northwest  21984.47061
4   32    male  28.880         0     no  northwest   3866.85520
```

[22]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

[23]:
```
df.describe()
```

[23]:
```
               age          bmi     children       charges
count  1338.000000  1338.000000  1338.000000   1338.000000
mean     39.207025    30.663397     1.094918  13270.422265
std      14.049960     6.098187     1.205493  12110.011237
min      18.000000    15.960000     0.000000   1121.873900
25%      27.000000    26.296250     0.000000   4740.287150
50%      39.000000    30.400000     1.000000   9382.033000
75%      51.000000    34.693750     2.000000  16639.912515
max      64.000000    53.130000     5.000000  63770.428010
```

[7]:
```

```

[8]:
```

```

[24]:
```
bool_series = pd.notnull(df['age'])
df[bool_series]
```

[24]:
```
      age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
1      18    male  33.770         1     no  southeast   1725.55230
2      28    male  33.000         3     no  southeast   4449.46200
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
…     …     …       …       …        …      …          …
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
1337   61  female  29.070         0    yes  northwest  29141.36030

[1338 rows x 7 columns]
```

[12]:
```
df.fillna('')
```

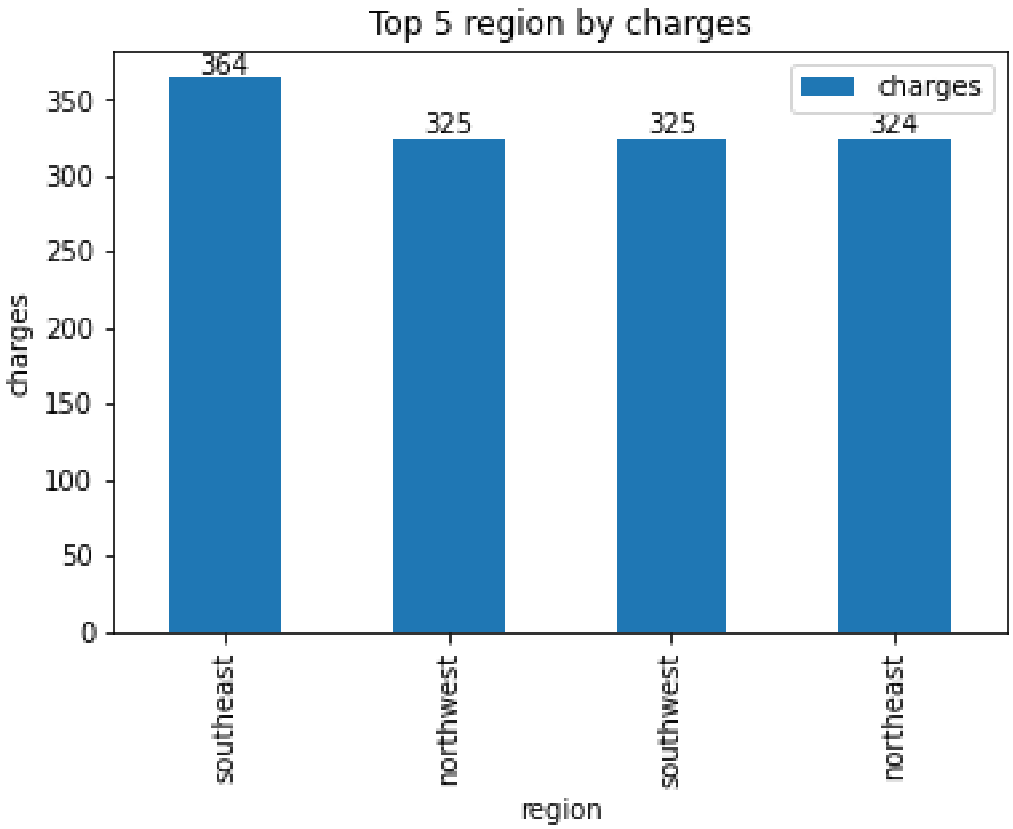
[12]:
```
      age     sex     bmi  children smoker     region      charges
0      19  female  27.900         0    yes  southwest  16884.92400
1      18    male  33.770         1     no  southeast   1725.55230
2      28    male  33.000         3     no  southeast   4449.46200
3      33    male  22.705         0     no  northwest  21984.47061
4      32    male  28.880         0     no  northwest   3866.85520
…     …     …       …       …        …      …          …
1333   50    male  30.970         3     no  northwest  10600.54830
1334   18  female  31.920         0     no  northeast   2205.98080
1335   18  female  36.850         0     no  southeast   1629.83350
1336   21  female  25.800         0     no  southwest   2007.94500
```

```
1337   61  female  29.070         0    yes  northwest  29141.36030

[1338 rows x 7 columns]
```

[15]:

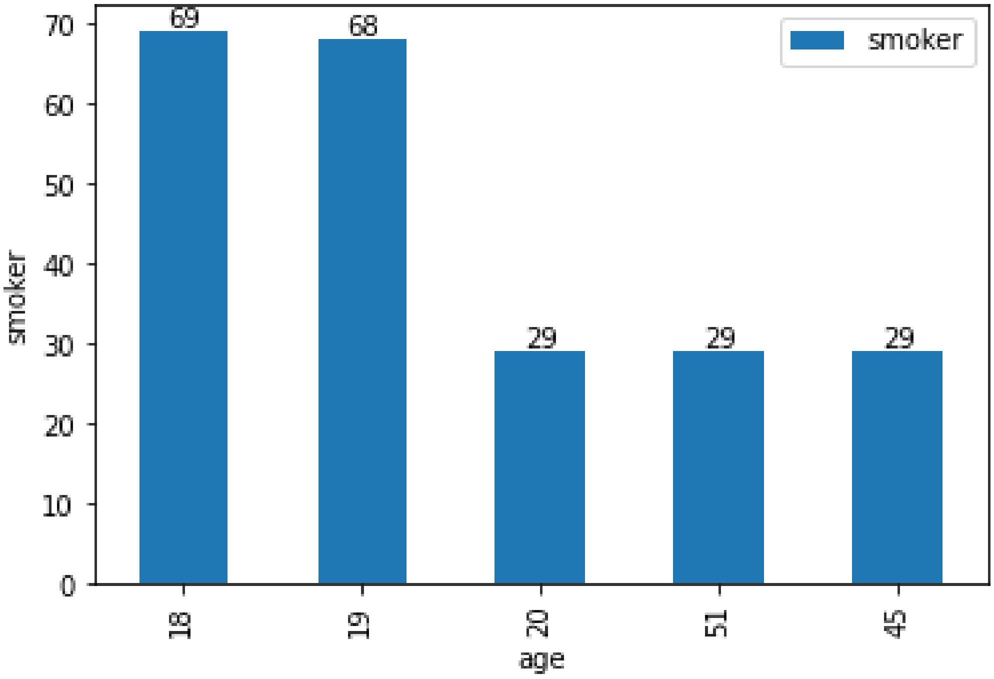
```
grp = df.groupby('region').agg({'charges':'count'})
z = grp.sort_values(by=['charges'],ascending=False)
ax = z.head().plot(kind='bar')
plt.xlabel('region')
plt.ylabel('charges')
plt.title('Top 5 region by charges')
for p in ax.patches:
  ax.annotate(str(p.get_height()),(p.get_x()+p.get_width()/2,p.
 ↪get_height()),ha='center',va ='bottom')
plt.show()
```



[25]:

```
grp = df.groupby('age').agg({'smoker':'count'})
z = grp.sort_values(by=['smoker'],ascending=False)
ax = z.head().plot(kind='bar')
plt.xlabel('age')
plt.ylabel('smoker')
```

```
for p in ax.patches:
  ax.annotate(str(p.get_height()),(p.get_x()+p.get_width()/2,p.
 ↪get_height()),ha='center',va ='bottom')
plt.show()
```



[26]:
```
df.shape
```

[26]:
```
(1338, 7)
```

[ ]: