U.S. Medical Insurance Costs

Figure out what the average age is for someone who has at least one child in this dataset.

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
df = pd.read_csv('insurance.csv')
```

In [2]:
```
df.head()
```

Out[2]:

	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

In [3]:
```
print(df.info())
print(df.describe())
```

```
<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
None
                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
```

In [26]:
```
genders_df = df.groupby("sex").agg([np.mean, np.std])
genders_df.head()
```

Out[26]:

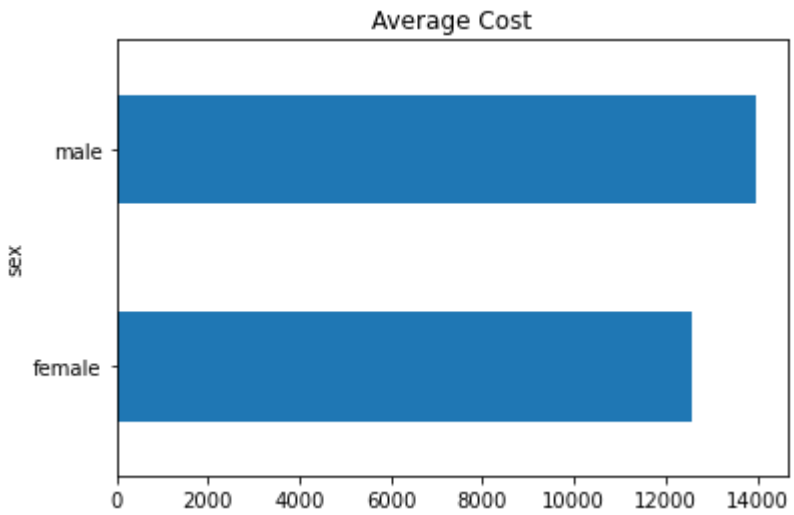
	age		bmi		children		charges	
	mean	std	mean	std	mean	std	mean	std
sex								
female	39.503021	14.054223	30.377749	6.046023	1.074018	1.192115	12569.578844	11128.703801
male	38.917160	14.050141	30.943129	6.140435	1.115385	1.218986	13956.751178	12971.025915

Here we can see that we do not have Null data and every cell is populated with data

This shows us the distribution of age by the count from this dataset

In [5]:
```
genders = genders_df['charges']
genders.plot(kind = "barh", y = "mean", legend = False,
             title = "Average Cost")
```

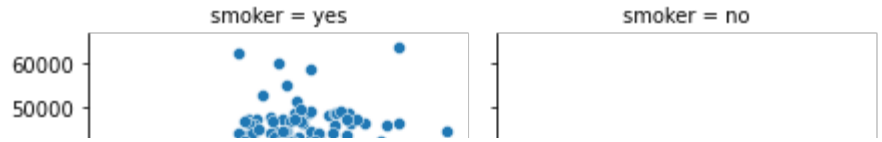
Out[5]: <AxesSubplot:title={'center':'Average Cost'}, ylabel='sex'>



From here we can see that the average cost of insurance is higher for men

In [19]:
```
grid = sns.FacetGrid(df, col = "smoker", hue = "smoker", col_wrap=5)
grid.map(sns.scatterplot, "bmi", "charges")
```

Out[19]: <seaborn.axisgrid.FacetGrid at 0x1db7d740c70>



This graph gives us a lot of insight about the cost of insurance based on bmi, and smoker status, as well as we can see the distribution between the bmi and smoker status

Now how about we try to find what would be the average cost of insurance for men with at least one child

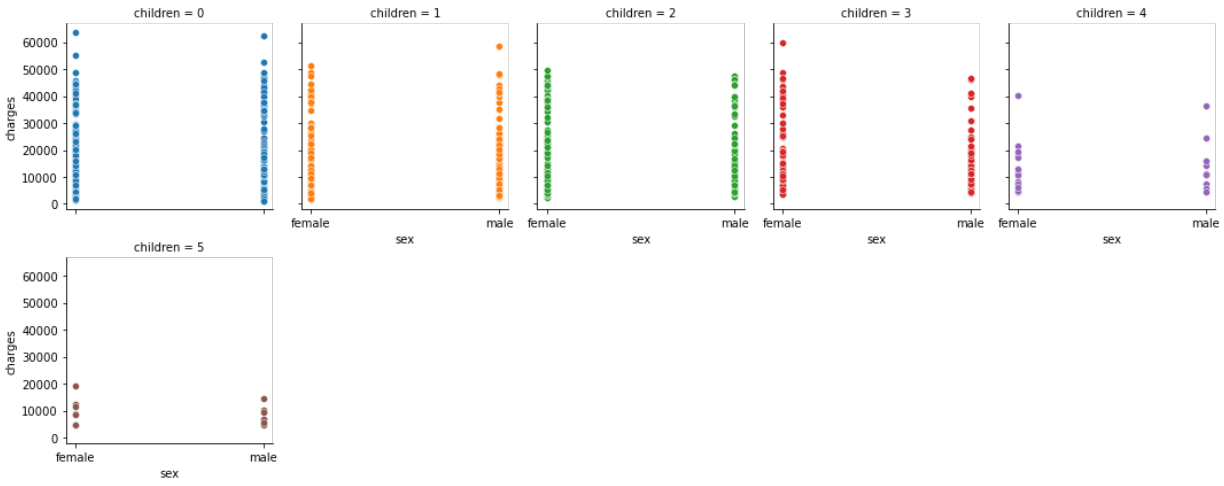
In [7]:
```
gender = df['sex']
children = df['children']
cost = df['charges']
gender_dict = {'male':[], 'female':[]}

for i in range(len(children)):
    if children[i] != 0:
        gender_dict[gender[i]].append((cost[i], children[i]))
    else:
        continue
avrg_cost_men_with_children = 0
for cost, num_children in gender_dict['male']:
    avrg_cost_men_with_children += cost
avrg_cost_men_with_children = avrg_cost_men_with_children / len(gender_dict['r
print('Average cost of insurance for men that have at least one child is ${}'
```

```
Average cost of insurance for men that have at least one child is $14776.07
```

In [25]:
```
grid = sns.FacetGrid(df, col = "children", hue = "children", col_wrap=5)
grid.map(sns.scatterplot, "sex", "charges")
```

Out[25]:
```
<seaborn.axisgrid.FacetGrid at 0x1db7f344610>
```



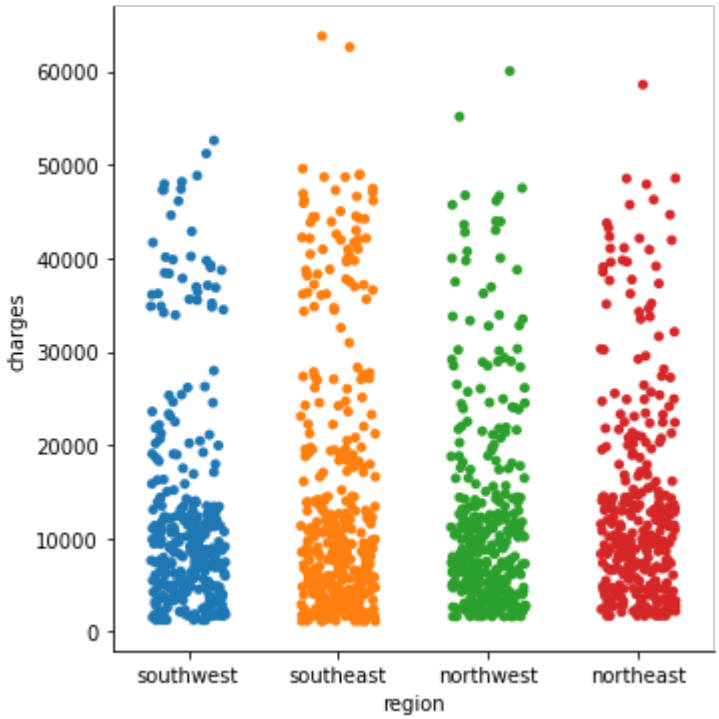
In [8]:
```
regions = df['region']
print('Here we can see the different regions in this dataset:')
for region in regions.unique():
    print(region)
```

```
Here we can see the different regions in this dataset:
southwest
southeast
northwest
northeast
```

In [27]:

```
sns.catplot(x='region', y='charges',
            data = df,
            jitter = '0.25')
print('Charges by the region')
```

```
Charges by the region
```

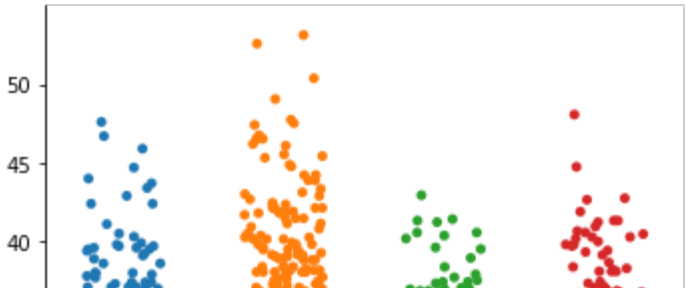


Here we can observe charges based on the regions and we can see that the distribution of values is not even, let's make a further analysis of this observation

In [14]:

```
sns.catplot(x='region', y='bmi',
            data = df,
            jitter = '0.25')
print('BMI by the region')
```

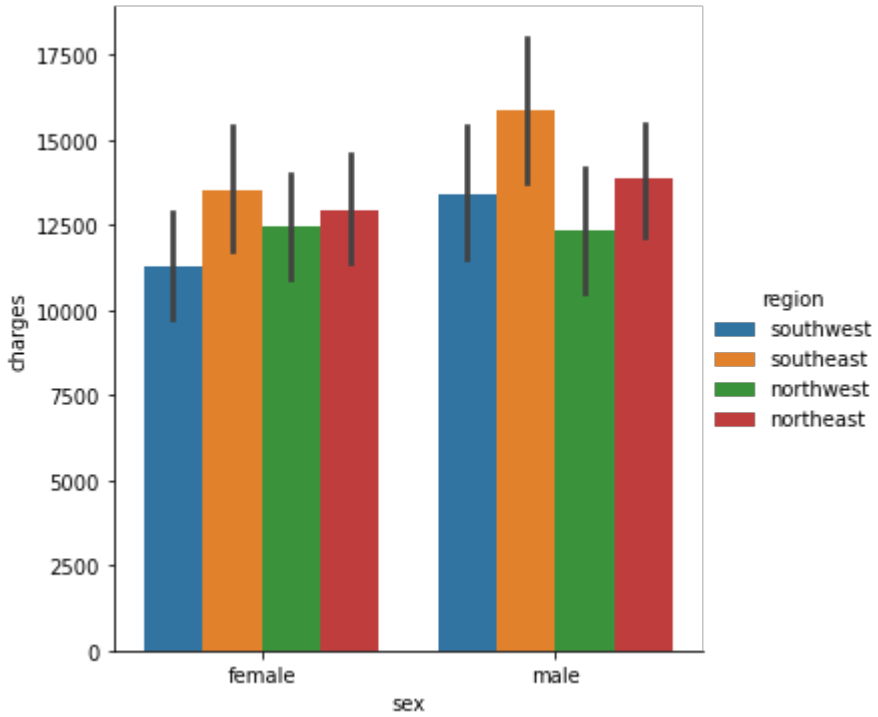
```
BMI by the region
```



Here we can see a better data regardging distribution of charges in different regions based on sex

In [11]:
```
sns.catplot(x="sex", y="charges", hue="region", kind="bar", data=df)
```

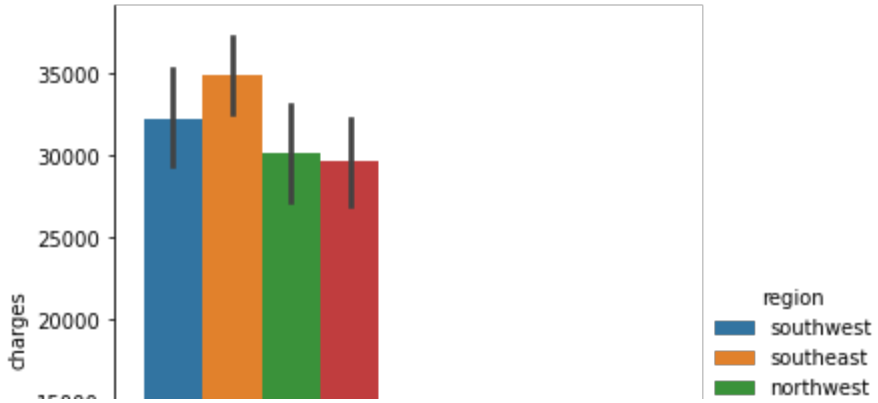
Out[11]: <seaborn.axisgrid.FacetGrid at 0x1db7d2619a0>



Here we can see a better data regardging distribution of charges in different regions based on smoking status

In [28]:
```
sns.catplot(x="smoker", y="charges", hue="region", kind="bar", data=df)
```

Out[28]: <seaborn.axisgrid.FacetGrid at 0x1db7fdd6880>


35000
30000
25000
20000
charges
region
southwest
southeast
northwest

In [ ]: