Cost of Living

Adeliia Salieva

Data Set and Variables

Data set is about the Cost of Living in almost 5000 cities across the world.

I chose this data set because it is interesting to analyze the cost of living in various places of the world.

There are 4,873 observations:

- Each observation is a cost of basic necessities and salary in cities across the world.

There are 13 variables. In this analysis we focus on the following ones:

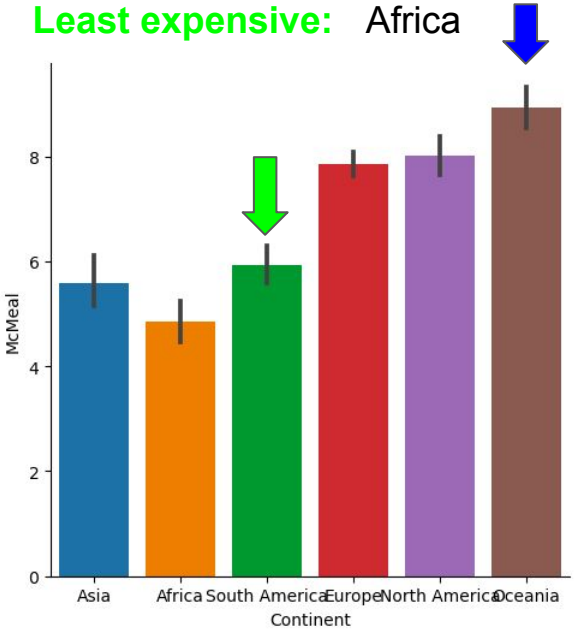
- City
- Country
- Continent
- McMeal (USD)
- Water (1.5 liter bottle) (USD)
- Gasoline (1 liter) (USD)
- Toyota Corolla Sedan (USD)
- Basic (Electricity, Heating, Cooling, Water, Garbage) for 85m2 Apartment (USD)
- Internet (USD)
- Preschool for 1 Child (USD)
- Jeans (USD)
- Apartment (1 bedroom) in City Centre (USD)
- Average Monthly Net Salary (After Tax) (USD)

Question 1: What are the costs of McMeal, Water and Gasoline in different Continents?

McMeal

Most expensive: Oceania

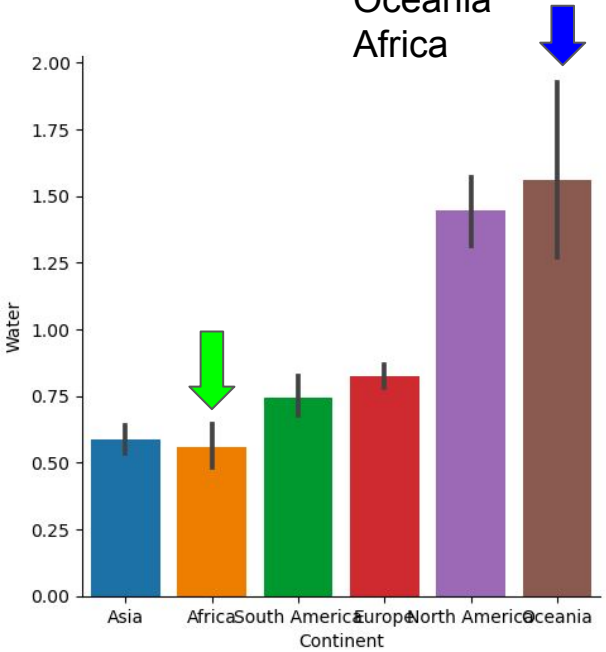
Least expensive: Africa



Water

Oceania

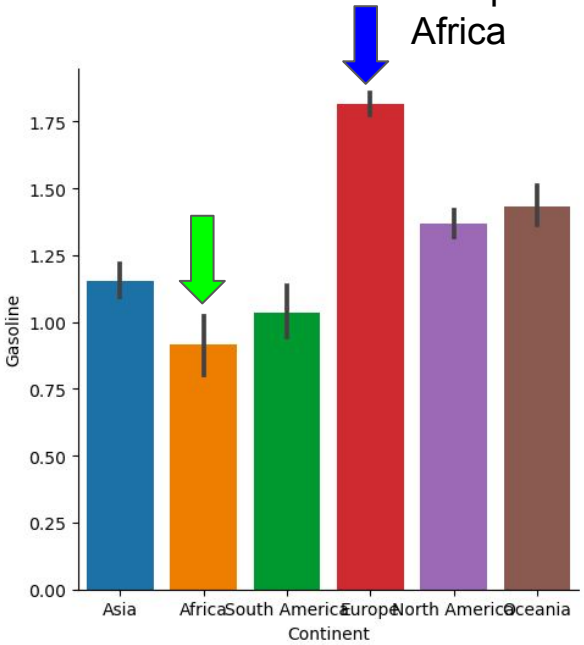
Africa



Gasoline

Europe

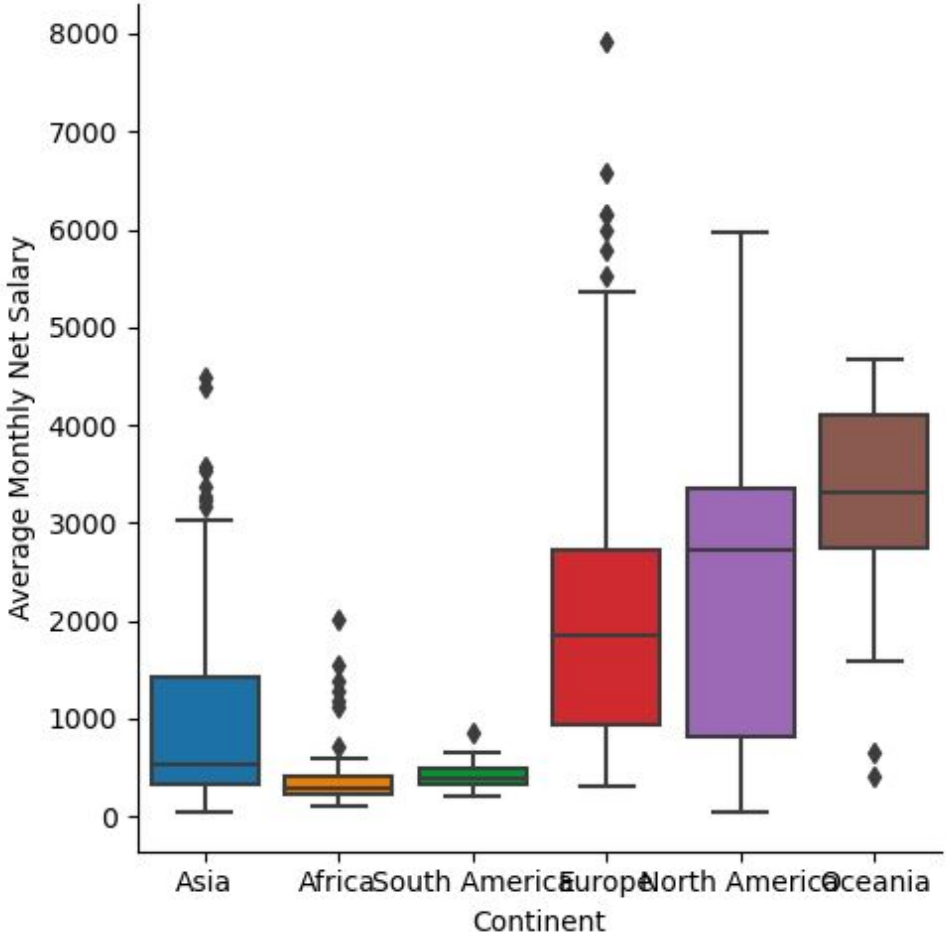
Africa



Question 2: What is the average Salary in different Continents?

Notes:

- The biggest average of Salary is in Oceania
- But the biggest ‘spreading’ of Salary is in Europe and North America



Hypothesis

Null: There is no difference in Salary between North America and Europe.

Alternative: There is a difference in Salary between North America and Europe.

T-test: Salary in North America vs Europe

```
# t-test
df_1 = df[df["Continent"] == 'North America']
df_2 = df[df["Continent"] == 'Europe']

stats.ttest_ind(df_1["Average Monthly Net Salary"], df_2["Average Monthly Net Salary"])
```

```
Ttest_indResult(statistic=1.931744582068827, pvalue=0.054018553041745l8)
```

> 0.05

What variables have the biggest influence on Salary variable?

Dependent Variable:

- Average Monthly Salary

Independent Variables:

- McMeal (USD)
- Water (1.5 liter bottle) (USD)
- Gasoline (1 liter) (USD)
- Toyota Corolla Sedan (USD)
- Basic (Electricity, Heating, Cooling, Water, Garbage) for 85m2 Apartment (USD)
- Internet (USD)
- Preschool for 1 Child (USD)
- Jeans (USD)
- Apartment (1 bedroom) in City Centre (USD)

Linear, Ridge, Lasso Regression models

	Linear	Ridge	Lasso
R^2 Score on train data:	0.7743	0.7743	0.7743
R^2 Score on test data:	0.7540	0.7544	0.7546

Linear

	variable	coefficient
3	Toyota	-0.007356
5	Internet	0.319839
8	Apt (1 bd) Centre	0.427274
7	Jeans	0.548882
4	Basic	0.856497
6	Preschool	1.042445
0	McMeal	118.272538
2	Gasoline	245.877957
1	Water	428.155966

Ridge

	variable	coefficient
3	Toyota	-0.007379
5	Internet	0.323010
8	Apt (1 bd) Centre	0.428703
7	Jeans	0.563230
4	Basic	0.863465
6	Preschool	1.042911
0	McMeal	118.786641
2	Gasoline	242.196161
1	Water	422.882327

Lasso

	variable	coefficient
3	Toyota	-0.007379
5	Internet	0.323010
8	Apt (1 bd) Centre	0.428703
7	Jeans	0.563230
4	Basic	0.863465
6	Preschool	1.042911
0	McMeal	118.786641
2	Gasoline	242.196161
1	Water	422.882327

Conclusions

1. There is no difference in Salary between North America and Europe.

2. Water, Gasoline and McMeal variables have the biggest influence on predicting Salary variable.

3. For improved generalization performance, it is better to use either the Ridge or Lasso model instead of the simple Linear model.

Thank you