

RQ2

Seungwon (Sydney) Lee Demi Leng Darlane Zhang
Hannah Chung

Thursday 18th December, 2025



1 Does affordability_ratio's mean differ across different regions, geotypes, and income levels?

In order to test this, I will be using a one-way ANOVA test. Since I have three different variables where I want to see if the different groups within the variables change the mean of affordability ratio, I'll be running the ANOVA 3 times. Then, I'll be using Tukey HSD to run post-hoc testing given we reject the null hypothesis within the ANOVA. Even though we have right-skewedness, ANOVA is still reasonable because of our large sample size and the robustness of ANOVA to deviations from normality. We assume that observations are independent and that variance is homogeneous.

For this test, my null hypothesis is that affordability_ratio's mean does not differ across different regions, different geotypes, and different income levels. Written out for region, geotype, and income level:

Region Null: The mean of affordability_ratio does not differ across different groups of region. Alternative: At least one mean of affordability_ratio differs from others across different groups of region.

Geotype Null: The mean of affordability_ratio does not differ across different groups of geotype. Alternative: At least one mean of affordability_ratio differs from others across different groups of geotype.

Income Level Null: The mean of affordability_ratio does not differ across different groups of income level. Alternative: At least one mean of affordability_ratio differs from others across different groups of income level.

```
import numpy as np
import pandas as pd
from pathlib import Path
import matplotlib.pyplot as plt
import seaborn as sns

from utils.model_utils import one_way_anova, tukey_hsd
```

```
RANDOM_STATE = 159
np.random.seed(RANDOM_STATE)

DATA_PATH = Path("./data/food_affordability.csv")
OUT_DIR = Path("./outputs"); OUT_DIR.mkdir(exist_ok=True)
FIG_DIR = Path("./figures"); FIG_DIR.mkdir(exist_ok=True)

df_2 = pd.read_csv(DATA_PATH)
df_2.head()
```

individually identifying the telephone number transferred to the individual's mobile phone number			
Individually Identifying the Telephone Number Transferred to the Individual's Mobile Phone Number			
0	757 Foo 2006.0 AIACA 6.0 CalNonNaN.. 237073052794004B99N4M285347	12/2013	4:33
	af- 2010 ford- abil- ity for female- headed house- hold...		
1	757 Foo 2006.0 AsiaGA 6.0 CalNonNaN.. 38508101980205895078101731940	12/2013	4:33
	af- 2010 ford- abil- ity for female- headed house- hold...		
2	757 Foo 2006.0 AfricaA 6.0 CalNonNaN.. 261028607900300B2733M0736044	12/2013	4:33
	af- 2010 ford- abil- ity for female- headed house- hold...		
3	757 Foo 2006.0 Latfia 6.0 CalNonNaN.. 2280830837263101297881233004	12/2013	4:33
	af- 2010 ford- abil- ity for female- headed house- hold...		
4	757 Foo 2006.0 NHORE 6.0 CalNonNaN.. 36787204379261097624255767348	12/2013	4:33
	af- 2010 ford- abil- ity for female- headed house- hold...		

```
5 rows × 23 columns
```

```
# Selecting only necessary columns
# I also dropped large outliers to reduce the skewedness slightly
rq2_cols = ["affordability_ratio", "median_income", "region_name", "geotype"]
df_rq2 = df_2[rq2_cols].copy()
df_rq2 = df_rq2.dropna(subset=["affordability_ratio", "median_income", "region_name", "geotype"])
upper = df_rq2["affordability_ratio"].quantile(0.99)
df_rq2 = df_rq2[df_rq2["affordability_ratio"] <= upper].copy()

df_rq2.shape

df_rq2["income_group"] = pd.qcut(
    df_rq2["median_income"],
    q=4,
    labels=["Q1 (Lowest)", "Q2", "Q3", "Q4 (Highest)"]
)

df_rq2["income_group"].value_counts()
df_rq2.groupby("income_group")["median_income"].mean()
df_rq2.groupby("income_group")["affordability_ratio"].mean()

income_group
Q1 (Lowest)      0.684877
Q2                0.303270
Q3                0.202455
Q4 (Highest)     0.116724
Name: affordability_ratio, dtype: float64
```

2 ANOVA for Regions

```
# ANOVA for region
anova_region = one_way_anova(df_rq2, target="affordability_ratio", group="region_name")
anova_region

{'F': 19.996761859495177,
 'p_value': 2.538578935644317e -49,
 'n_groups': 15,
 'n_total': 3439}
```

My F-statistic for this test was 19.996, which leads me to a p-value of 2.54 times e^{-49} . This very small p-value leads me to reject the null hypothesis and we conclude that there is statistically significant evidence that at least one regional mean affordability ratio differs from the others.

Now, I will run post-hoc testing with Tukey HSD.

```
# Tukey for region
tukey_region = tukey_hsd(df_rq2, "affordability_ratio", "region_name")
tukey_region.head()
```

	group1	group2	meandiff	p-adj	lower	upper	reject
0	Bay Area	Butte	0.1848	0.0066	0.0266	0.3430	True
1	Bay Area	California	0.0164	1.0000	-0.3487	0.3815	False
2	Bay Area	Central/Southeast	0.0663	0.0663	-0.0036	0.2633	False
3	Bay Area	Monterey Bay	0.1557	0.0000	0.0546	0.2568	True
4	Bay Area	North Coast	0.1623	0.0000	0.0542	0.2704	True

Tukey HSD results indicate that the Bay Area exhibits significantly higher food affordability ratio than other inland regions like Butte, Monterey Bay, and North Coast. The reason we know that Bay Area food affordability ratio is larger is because the mean differences are positives with a low p-val. This shows a greater food cost burden in coastal regions.

3 ANOVA for Geotype

```
# ANOVA for geotype
anova_geotype = one_way_anova(df_rq2, target="affordability_ratio", group="geotype")
anova_geotype

{'F': 0.2205838707762331,
 'p_value': 0.8821537113878416,
 'n_groups': 4,
 'n_total': 3439}
```

My F-statistic for this test was 0.22, which leads me to a p-value of 0.88. This large p-value leads me to fail to reject the null hypothesis and we conclude that there is no statistically significant evidence that at least one regional mean affordability ratio differs from the others.

I will not run post-hoc testing with Tukey HSD in this case.

4 ANOVA for Income Level

```
# ANOVA for income level
anova_income_level = one_way_anova(df_rq2, target="affordability_ratio", group="income_group")
anova_income_level
```

```

{'F': 850.1216693204053, 'p_value': 0.0, 'n_groups': 4, 'n_total': 3439}

# Tukey for income level
tukey_income_level = tukey_hsd(df_rq2, "affordability_ratio", "income_group")
tukey_income_level.head()

```

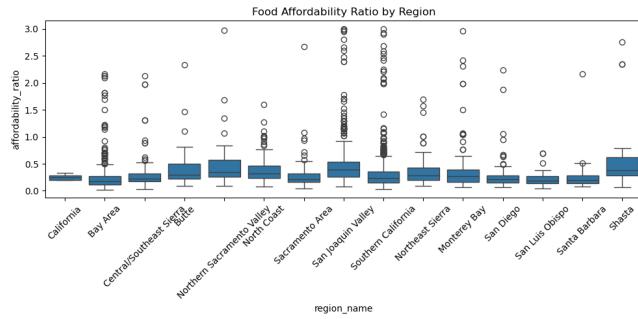
	group1	group2	meandiff	p-adj	lower	upper	reject
0	Q1 (Low- est)	Q2	-0.3816	0.0	-0.4129	-0.3504	True
1	Q1 (Low- est)	Q3	-0.4824	0.0	-0.5137	-0.4512	True
2	Q1 (Low- est)	Q4 (High- est)	-0.5682	0.0	-0.5994	-0.5369	True
3	Q2	Q3	-0.1008	0.0	-0.1321	-0.0695	True
4	Q2	Q4 (High- est)	-0.1865	0.0	-0.2178	-0.1553	True

Post-hoc with Tukey HSD shows there are statistically significant differences in mean food affordability ratios throughout Q1, Q2, Q3, and Q4. Even though affordability ratios increase as you go from Q1 to Q4, every higher income group shows a larger food cost burden than the group before it. The higher income regions show higher affordability ratios, which might suggest that food costs are disproportionately increasing with income level across Californian regions.

```

plt.figure(figsize=(10, 5))
sns.boxplot(
    data=df_rq2,
    x="region_name",
    y="affordability_ratio"
)
plt.xticks(rotation=45)
plt.title("Food Affordability Ratio by Region")
plt.tight_layout()
plt.savefig('figures/rq2_boxplot.png', dpi=150)
plt.show()

```



This figure shows the spread of food affordability ratio by region.

5 Conclusion for RQ2

For this question, I aimed to look at potential differences in mean of affordability ratio in region, income level, and geotype. I concluded that there were significant differences in mean between region and income level, but not geotype.