Hypothesis Testing

Rahul Sharma
500091839

R2142210619

B.Tech. CSE AIML Batch-2 Non-hons.

Applying Z-Test T-Test and ANOVA on house price dataset.

Got the dataset from Kaggle.
The real estate markets, like those in Sydney and Melbourne, present an interesting opportunity for data analysts to analyze and predict where property prices are moving towards. Prediction of property prices is becoming increasingly important and beneficial. Property prices are a good indicator of both the overall market condition and the economic health of a country. Considering the data provided, we are wrangling a large set of property sales records stored in an unknown format and with unknown data quality issues.

```
# Step 1: T-Tests for 'sqft_lot' and 'bedrooms'
X_sqft_lot = df['sqft_lot']
X_bedrooms = df['bedrooms']
y = df['price']
```

[10] ✓ 0.0s Python

+ Code + Markdown

```
# T-Test for 'sqft_lot'
t_stat_sqft_lot, p_val_sqft_lot = stats.ttest_ind(X_sqft_lot, y)
print("T-Test for sqft_lot:")
print("T-Statistic:", t_stat_sqft_lot)
print("P-Value:", p_val_sqft_lot)
```

[11] ✓ 0.0s Python

```
T-Test for sqft_lot:
T-Statistic: -64.47820750459283
P-Value: 0.0
```

Conclusion:

With a p-value effectively at 0.0, it suggests strong evidence against the null hypothesis. Here's how you'd interpret this result:

- **Conclusion**:
  - We reject the null hypothesis. There is strong evidence to suggest that `sqft_lot` has a statistically significant effect on the `Property_Price`.
- **In Practical Terms**:
  - The negative t-statistic indicates that, on average, as `sqft_lot` increases, the `Property_Price` tends to decrease. This suggests a relationship where larger lot sizes might lead to lower property prices.

```
# T-Test for 'bedrooms'
t_stat_bedrooms, p_val_bedrooms = stats.ttest_ind(X_bedrooms, y)
print("\nT-Test for bedrooms:")
print("T-Statistic:", t_stat_bedrooms)
print("P-Value:", p_val_bedrooms)
```

[12] ✓ 0.0s Python

```
T-Test for bedrooms:
T-Statistic: -66.39485018501195
P-Value: 0.0
```

Conclusion:

With a p-value effectively at 0.0, it suggests strong evidence against the null hypothesis. Here's how you'd interpret this result:

- **Conclusion**:
  - We reject the null hypothesis. There is strong evidence to suggest that the `bedrooms` variable has a statistically significant effect on `Property_Price`.
- **In Practical Terms**:
  - The negative t-statistic indicates that, on average, as the number of `bedrooms` increases, the `Property_Price` tends to decrease. This suggests a relationship where properties with more bedrooms might have lower prices.

```python
# Step 2: ANOVA for 'Location'
model = sm.formula.ols('price ~ city', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)

print("\nANOVA for Location:")
print(anova_table)
```

[14] ✓ 0.0s Python

```
ANOVA for Location:
                sum_sq      df          F        PR(>F)
city      1.571925e+14    43.0  12.763756  2.990824e-83
Residual  1.304874e+15  4556.0        NaN           NaN
```

Conclusion:

Based on the ANOVA results:

- **Conclusion**:
  - We reject the null hypothesis. There is strong evidence to suggest that the `Location` (`city`) variable has a statistically significant effect on `Property_Price`.
- **In Practical Terms**:
  - This means that the `Location` (city) where a property is located is likely to have a significant impact on its `Property_Price`. Different cities tend to have different average property prices.