ANOVA

In [22]:
```
import pandas as pd
import scipy.stats as stats
import statsmodels.api as sm
from statsmodels.formula.api import ols
import scikit_posthocs as sp
```

In [4]:
```
data_anova = pd.read_excel('/Users/yashgupta/Downloads/BSAN Project Files Edited/Final Datasets/Alcohol and Drug Abuse Subset.xlsx')
```

In [5]:
```
data_anova.isna().sum()
```

Out[5]:
```
HOSP_REGION_DES    0
RACE_DES           0
TOTCHG             0
dtype: int64
```

In [6]:
```
# Check for non-numeric values in the 'TOTCHG' column
non_numeric = data_anova['TOTCHG'].apply(lambda x: isinstance(x, str))
missing_data = data_anova['TOTCHG'].isnull()

# Remove rows with non-numeric 'TOTCHG' values
clean_data_anova = data_anova[~(non_numeric | missing_data)].copy()

# Convert 'TOTCHG' to numeric
clean_data_anova['TOTCHG'] = pd.to_numeric(clean_data_anova['TOTCHG'])

# Perform the Two-Way ANOVA
model_anova_clean = ols('TOTCHG ~ C(HOSP_REGION_DES) + C(RACE_DES) + C(HOSP_REGION_DES):C(RACE_DES)', data=clean_data_anova).fit()
anova_results_clean = sm.stats.anova_lm(model_anova_clean, typ=2)

# Output the results
print(anova_results_clean)
```

```
                                  sum_sq      df          F        PR(>F)
C(HOSP_REGION_DES)            8.250780e+11     3.0  78.920493  4.357724e-50
C(RACE_DES)                   2.184381e+11     5.0  12.536444  3.914456e-12
C(HOSP_REGION_DES):C(RACE_DES)  1.026040e+11    15.0   1.962858  1.426386e-02
Residual                      2.134819e+13  6126.0        NaN           NaN
```

Kruskal-Wallis

In [14]:
```
group_sizes = clean_data_anova.groupby(['HOSP_REGION_DES', 'RACE_DES']).size()
print(group_sizes)
```

```
HOSP_REGION_DES  RACE_DES
Midwest          Asian or Pacific Islander      23
                 Black                         232
                 Hispanic                      105
                 Native American                57
                 Other                          59
                 White                        1154
Northeast        Asian or Pacific Islander      27
                 Black                         167
                 Hispanic                      193
                 Native American                 2
                 Other                         116
                 White                         493
South            Asian or Pacific Islander      17
                 Black                         458
                 Hispanic                      332
                 Native American                16
                 Other                          93
                 White                        1366
West             Asian or Pacific Islander      73
                 Black                          82
                 Hispanic                      397
                 Native American                59
                 Other                          76
                 White                         553
dtype: int64
```

In [16]:
```
# Filter groups with at least 3 observations
filtered_data = clean_data_anova.groupby(['HOSP_REGION_DES', 'RACE_DES']).filter(lambda x: len(x) >= 3)
```

In [18]:
```
# Group by each factor and check if there are at least two groups
groups_by_region = filtered_data.groupby('HOSP_REGION_DES')
groups_by_race = filtered_data.groupby('RACE_DES')

# Kruskal-Wallis Test for 'HOSP_REGION_DES' if there are at least two groups
if len(groups_by_region) >= 2:
    kruskal_results_region = stats.kruskal(*[group['TOTCHG'] for name, group in groups_by_region])
    print('Kruskal-Wallis Test for HOSP_REGION_DES:', kruskal_results_region)
else:
    print('Not enough groups for Kruskal-Wallis Test on HOSP_REGION_DES')

# Kruskal-Wallis Test for 'RACE_DES' if there are at least two groups
if len(groups_by_race) >= 2:
    kruskal_results_race = stats.kruskal(*[group['TOTCHG'] for name, group in groups_by_race])
    print('Kruskal-Wallis Test for RACE_DES:', kruskal_results_race)
else:
    print('Not enough groups for Kruskal-Wallis Test on RACE_DES')
```

```
Kruskal-Wallis Test for HOSP_REGION_DES: KruskalResult(statistic=413.96141641337977, pvalue=2.0935623622457854e-89)
Kruskal-Wallis Test for RACE_DES: KruskalResult(statistic=204.29404777163893, pvalue=3.4251160817379375e-42)
```

In [23]:
```
# Dunn's Test for 'HOSP_REGION_DES'
dunn_test_region = sp.posthoc_dunn(filtered_data, val_col='TOTCHG', group_col='HOSP_REGION_DES', p_adjust='bonferroni')
print('Dunn\'s Test for HOSP_REGION_DES:\n', dunn_test_region)

# Dunn's Test for 'RACE_DES'
dunn_test_race = sp.posthoc_dunn(filtered_data, val_col='TOTCHG', group_col='RACE_DES', p_adjust='bonferroni')
print('Dunn\'s Test for RACE_DES:\n', dunn_test_race)
```

```
Dunn's Test for HOSP_REGION_DES:
                Midwest     Northeast         South          West
Midwest    1.000000e+00  2.442636e-50  2.389712e-14  3.770388e-71
Northeast  2.442636e-50  1.000000e+00  1.805234e-19  5.715719e-01
South      2.389712e-14  1.805234e-19  1.000000e+00  4.644086e-32
West       3.770388e-71  5.715719e-01  4.644086e-32  1.000000e+00
Dunn's Test for RACE_DES:
                           Asian or Pacific Islander         Black  \
Asian or Pacific Islander                1.000000e+00  1.000000e+00
Black                                    1.000000e+00  1.000000e+00
Hispanic                                 1.000000e+00  1.967826e-02
Native American                          7.042033e-08  4.744021e-09
Other                                    1.000000e+00  4.865932e-02
White                                    3.572575e-04  1.258790e-09

                             Hispanic  Native American         Other  \
Asian or Pacific Islander  1.000000e+00     7.042033e-08  1.000000e+00
Black                      1.967826e-02     4.744021e-09  4.865932e-02
Hispanic                   1.000000e+00     4.417334e-14  1.000000e+00
Native American            4.417334e-14     1.000000e+00  8.019632e-13
Other                      1.000000e+00     8.019632e-13  1.000000e+00
White                      4.042140e-26     1.384044e-03  9.237443e-13

                                 White
Asian or Pacific Islander  3.572575e-04
Black                      1.258790e-09
Hispanic                   4.042140e-26
Native American            1.384044e-03
Other                      9.237443e-13
White                      1.000000e+00
```

These results from the Kruskal-Wallis tests for 'HOSP_REGION_DES' and 'RACE_DES' show significant differences in the 'TOTCHG' variable across the different levels of these categorical variables. Let's interpret these results:

Kruskal-Wallis Test for 'HOSP_REGION_DES':

Statistic: 413.96413.96 P-value: 2.09 \times 10^{-89}2.09×10 −89

Interpretation: This extremely low p-value suggests that there are significant differences in the 'TOTCHG' values across the different hospital regions. The high test statistic value further indicates strong evidence against the null hypothesis of identical distributions of 'TOTCHG' across different hospital regions.

Kruskal-Wallis Test for 'RACE_DES':

Statistic: 204.29204.29 P-value: 3.43 \times 10^{-42}3.43×10 −42

Interpretation: Similarly, this result indicates significant differences in the 'TOTCHG' values across different racial groups. The low p-value rejects the null hypothesis, suggesting that at least one racial group has a different distribution of 'TOTCHG' compared to others.

Interpretation of Dunn's Test Results:

For 'HOSP_REGION_DES':

The p-values are shown for each pair of regions. A p-value less than 0.05 typically indicates a statistically significant difference. For example, the p-value between Midwest and Northeast is approximately 2.44 \times 10^{-50}2.44×10 −50 , indicating a statistically significant difference in 'TOTCHG' between these two regions. Similarly, significant differences are observed between several other pairs of regions, as indicated by the very low p-values (e.g., Midwest and West, South and West).

For 'RACE_DES':

The table shows p-values for pairwise comparisons between different racial groups. Many pairs show significant differences. For instance, the p-value between Hispanic and Native American is about 4.42 \times 10^{-14}4.42×10 −14 , suggesting a significant difference in 'TOTCHG' between these two racial groups. However, some comparisons do not show significant differences (p-values close to 1), such as between Asian or Pacific Islander and Black.

In [ ]:
```
```